DOCUMENT RESUME

ED 161 846 95 SP 013 226

AUTHOR Fennessey, James; Horiuchi, Shiro
TITLE An Experiment on Social Information Processing Among Teachers. Report No. 243.
INSTITUTION Johns Hopkins Univ., Baltimore, Md. Center for Social Organization of Schools.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Jan 78
CONTRACT NE-C-00-3-0114
NOTE 105p.

EDRS PRICE MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS Data Analysis; *Educational Experiments; Educational Improvement; Elementary Secondary Education; *Evaluation Methods; *Information Processing; *Management Information Systems; Measurement Techniques; Operations Research; *School Administration; *Teacher Participation

ABSTRACT

School management activities can be improved by the more complete and systematic use of information on the concrete situation. Teacher judgments constitute an underutilized source of such information, e. g., concerning the distribution of student achievement levels. In this study, the application of formal techniques from management science (probability assessment techniques and structured group interaction) is investigated as a means of obtaining such information in a convenient quantitative form. Using a variation on a Latin Square design, several possible influences upon teacher judgments made in response to a realistic scenario are evaluated. The results generally indicate that application of such techniques is feasible and useful. The teachers' judgments are not greatly affected by the particular instrument format used to obtain their assessments, nor by the sequence in which particular items of information are received. The experimental data also provide evidence that group interaction among the teacher-judges tends to strengthen the influence of their prior beliefs. Other experimental results suggest that in general the teachers are not "conservative" users of information, but that conservatism may occur to some degree when the received information is surprising. An additional finding of interest is that the aggregate judgments of the teachers are extremely accurate and nicely responsive to new information. However, this is partly counterbalanced by the occurrence of sizeable individual variation between teachers, so that an individual teacher's judgments are not nearly so valuable as those of a small group. (Author/DS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for Social Organization of Schools

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED161846

REPORT NO. 243
JANUARY 1978

AN EXPERIMENT ON SOCIAL INFORMATION PROCESSING AMONG TEACHERS

JAMES FENNESSEY AND SHIRO HORIUCHI

The Johns Hopkins University

SP013226

STAFF

Edward L. McDill, Co-director

James M. McPartland, Co-director

Karl L. Alexander

Henry J. Becker

Bernard L. Blackburn

Vicky C. Brown

Martha A. Cook

Denise C. Daiger

Joyce L. Epstein

James J. Fennessey

Gary D. Gottfredson

Linda S. Gottfredson

Larry J. Griffin

Edward J. Harsch

John H. Hollifield

Lawrence F. Howe

Barbara J. Hucksoll

Nancy L. Karweit

Hazel G. Kennedy

Marshall B. Leavey

Anne McLaren

James M. Richards, Jr.

Margaret Ann Ricks

Richard R. Scott

Robert E. Slavin

Charles B. Thomas

Gail E. Thomas

An Experiment on Social Information Processing Among Teachers

CONTRACT NO. NE-C-00-3-0114

James Fennessey

Shiro Horiuchi

Report No. 243
January 1978

Published by the Center for Social Organization of Schools, supported in part as a research and development center by funds from the United States National Institute of Education, Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the Institute should be inferred.

The Johns Hopkins University

Baltimore, Maryland

Introductory Statement

The Center for Social Organization of Schools has two primary objectives: to develop a scientific knowledge of how schools affect their students, and to use this knowledge to develop better school practices and organization.

The center works through three programs to achieve its objectives. The Policy Studies in School Desegregation program applies the basic theories of social organization of schools to study the internal conditions of desegregated schools, the feasibility of alternative desegregation policies, and the interrelation of school desegregation with other equity issues such as housing and job desegregation. The School Organization program is currently concerned with authority-control structures, task structures, reward systems, and peer group processes in schools. It has produced a large-scale study of the effects of open schools, has developed the Teams-Games-Tournament (TGT) instructional process for teaching various subjects in elementary and secondary schools, and has produced a computerized system for school-wide attendance monitoring. The School Process and Career Development program is studying transitions from high school to post secondary institutions and the role of schooling in the development of career plans and the actualization of labor market outcomes.

This report, prepared by the School Organization Program, presents an experiment designed to investigate the social dynamics and outcomes of teachers working together to process information and make evaluations.

ABSTRACT

School management activities can be improved by the more complete and systematic use of information on the concrete situation. Teacher judgments constitute one underutilized source of such information--for example, concerning the distribution of student achievement levels. In this study, we investigated the application of formal techniques from management science (probability assessment techniques and structured group interaction) as a means of obtaining such information in a convenient quantitative form. Using a variation on a Latin Square design, we evaluated several possible influences upon teacher judgments made in response to a realistic scenario. The results generally indicate that application of such techniques is feasible and useful. The teachers' judgments are not greatly affected by the particular instrument format used to obtain their assessments, nor by the sequence in which particular items of information are received. The experimental data also provide evidence that group interaction among the teacher-judges tends to strengthen the influence of their prior beliefs. Other experimental results suggest that in general the teachers are not "conservative" users of information, but that conservatism may occur to some degree when the received information is surprising. An additional finding of interest is that the aggregate judgments of the teachers are extremely accurate and nicely responsive to new information. However, this is partly counterbalanced by the occurrence of sizeable individual variation between teachers, so that an individual teacher's judgments are not nearly so valuable as those of a small group. The report concludes with some suggestions for further research.

ACKNOWLEDGMENTS

This Curriculum Assessment research project was greatly facilitated by the interest and cooperation of Wilmer L. Jones, Jr., of the Baltimore City Public Schools. We appreciate his assistance. Thanks are due also to the group of teachers, numbering nearly 50 persons, who participated in the project as expert judges. Their alertness and professionalism impressed all of us who worked on the project.

During the planning and data collection phases, we were fortunate to have several dedicated and competent colleagues as co-workers. Special thanks are in order to Joel Gelb, Joanne Julius, Kelly K'Meyer, and Anne Wilson, who provided helpful ideas and suggestions in addition to their other activities on the project.

Naturally, although we have benefited from the input provided by all of these persons, and by other colleagues at the Center for Social Organization of Schools and the Social Relations Department, we must as authors assume sole responsibility for any errors or defects of this report.

J.F.
S.H.

1. Introduction

Over the past few years, several studies have investigated the benefits of using quantitative data inputs, and formalized operating systems, in school management (Fennessey, 1973a, 1973b; Karweit, 1973; Richards, 1976). These studies, however, were concerned primarily with delivering information to school administrators, or with the statistical analysis of achievement test results. There has been little investigation of the social dynamics and outcomes that are likely when staff members in a school work together to develop or modify school procedures, taking quantitative information into account systematically. Similarly, despite a growing interest in the individual teacher as a classroom manager and clinical problem-solver (cf. Shulman and Elstein, 1975), little attention has been given to the overall school management process as a patterned social activity, one in which teachers and administrators work together collegially to define or revise the school's operating style.

2. Study Design Considerations

For several reasons, our study concentrates on information processing in school management contexts, rather than on personal interests, values, or ideology. Clearly, just as in corporations and government agencies, actual decisions and policies do depend upon participants' personal values, and on political factors as well. However, management science techniques for dealing with these latter aspects are less advanced than those for handling information inputs. Also, changes in the information processing procedures of school management should be less controversial,

and thus provide a more acceptable first step in updating school operating practices. Finally, there is a research literature--principally laboratory studies in psychology and applied experiments in medical diagnosis--on the nature of clinical information processing by individuals and groups. That literature, though not directly applicable to school operations, does offer some guidance as to possible issues and patterns to be investigated.

Thus, in this project we study how teachers, individually and in groups, can use different kinds of evidence in forming a judgment about some facet of the school situation. Achievement levels and achievement test results are one important and familiar concern of school management activities, and mathematics as a subject area is a frequent focus of attention. We decided to work with junior high school mathematics teachers, and to study how different procedures and circumstances affect the judgments they make about the distribution of mathematics skills in a particular group of students. Mathematics was chosen as a subject area because (a) it lends itself well to measurement by test scores, and (b) mathematics teachers, compared to teachers of other subjects, are likely to be more comfortable with quantitative ideas and thinking.

Another study design question concerns the degree of realism in the situation to be presented to the teacher-judges. At one extreme, the teachers might be given purely hypothetical test information and asked to make their judgments. This would provide a clear set of results, in that few extraneous stimuli would have been introduced. At the other extreme, several teachers from a single school might be assembled and asked for judgments about the mathematics skills in that school. Such an approach

would make maximal use of the teachers' actual expert knowledge of the situation, but each group of teachers would be non-comparable to any other group, and the stimuli they were judging would likewuse be non-comparable. Also, such discussion and assessments of their own school might have undesirable impacts on the teachers in their work thereafter. Rather than choose either of these extreme possibilities, we decided to use a scenario involving a fictitious school, but providing considerable information about that school and its surrounding community. This strategy obtains most of the benefits of using the real school situation, maintains comparability of stimuli across groups, allows comparison of separate groups, and precludes feedback into any actual schools.

Development and implementation of this strategy was greatly facilitated by our use of the FEHR-Practicum Package (Collet, 1974). This package, comprising a set of computer programs, reference materials, and user's manuals, was originally designed as a tool for training educational researchers. One segment of it, the "Remedial Arithmetic" packet, provides written materials describing an urban community and one of its junior high schools. The materials also outline a curriculum planning challenge to be addressed. The corresponding computer programs of FEHR provide, on request, various data for the students in the school, including scores on the mathematics achievement subtest of the Iowa Tests of Basic Skills. These FEHR data have been painstakingly constructed so that they are representative of the best available knowledge about school achievement patterns. Using these materials, with some adaptations and modifications, we created a reasonable, "high fidelity" scenario

with which small groups of teachers could form concrete judgments about the patterns of mathematics skills in a school, and could react to various kinds of information input.

The basic idea behind the management science techniques investigated in this study is a simple one. Most briefly, it is that many phenomena of interest (such as mathematics achievement) take the form of distributions of scores within some population. These distributions cannot be fully described by any single number, or even by any small set of numbers. They may, however, be fully described by a plotted curve or graph. In such graphs, the horizontal axis represents the range of possible scores, and the vertical axis represents some measure of relative occurrence or cumulative occurrence.

Such graphs can represent equally well either of two distinct concepts. They may be used to represent a distribution of actual values in some specified population (as we use them in this project), or they may be used to represent a person's beliefs about the likelihood or probability of different possible scores occurring in one particular case. This latter usage is compatible with the so-called Bayesian paradigm of inference and decision analysis (cf. Fennessey, 1977). It provides a new and powerful means of formalizing clinical judgment. Such judgment is, after all, simply a careful application of general knowledge about a phenomenon together with specific facts about one case, to draw a probabilistic conclusion or assessment of that case. Thus, our work here with distributional judgments concerning groups of students has relevance as well for the judgments teachers make about individual students.

We conceptualized the judgment formation process as having two main aspects. The first aspect is the attempt to identify relevant predictors, or indicators, or clues that should be sought and considered ("What are the symptoms to be checked?"). This first aspect seeks to analyze or decompose the phenomenon and identify relevant specific features. We have called it "diagnosis." The second aspect of the process consists of taking all the available clues --however incomplete--and forming a conclusion. In cases where that conclusion is a quantitative estimate (including our study) we call this synthesizing aspect "assessment."

3. Research Aims

In this study, our aims were several. First, we sought to ascertain the sensitivity of the assessments actually made by teachers under different conditions. One such condition is simply the format of the assessment instrument. To examine this, half of the teachers were asked to provide their assessments using density curves, and the other half were asked to use cumulative curves. A second interest was the sensitivity of assessments to the perceived source of information. Thus, some of the information was described as coming from judgments by other teachers, and some of it as coming from objective tests given to a sample of the students. A third interest was whether information indicating high achievement is responded to differently than is information indicating low achievement.

In addition to these three interests--each concerned with the information processing activity by individual teachers--we were interested also in the differences in assessment, if any, when group interaction

among the teachers occurred relatively early versus relatively late. Finally, we sought to explore and elucidate the diagnosis process, both individual and social. That is, we sought to learn what factors were chosen by the teachers as relevant, and how the likely importance and actual values of these factors were perceived in a concrete situation.

With respect to the revision of judgments in light of new information, there is a formal rule (Bayes Theorem) which prescribes the optimal amount of such revision, given certain amounts of information. An interesting question, but one seldom studied in realistic settings, and never with teachers, is how human experts (either individually or in groups) behave in relation to this theorem. Do they use information optimally? Some studies in the psychology laboratory (cf. Edwards, 1968) tend to indicate that people are conservative; they attach too little import to the evidence they receive. This line of argument would be consistent with the idea sometimes heard that teachers rely heavily upon stereotypes, and downgrade any new information that may be presented. Thus, the responsiveness of teachers after receiving new information was an additional question of interest in the present study.

The foregoing paragraphs make clear that our aims were multiple. In part we hoped for some clear experimental comparisons, but equally important was the exploratory nature of the study and the procedural experience to be acquired by carrying it out. Many of the experimental procedures had not been even approximately tried prior to our use of them. Thus, the present study also serves as a field trial of a new experimental approach to the study of judgment formation.

4. Design of the Experiment

Our study proceeded in two separate phases. These phases concentrated on different issues and were separated in time by several months. The first phase was designed to ascertain the impacts of several of the factors already described on the actual assessments (i.e., on the quantified judgments as to the distribution of mathematics skills) made by teacher-experts in a concrete situation. Data on the diagnosis portion of the judgment formation process (i.e., thinking about possible clues and their relevance to the instance being judged) was also collected, but was secondary in importance. The experiment was designed to allow statistical comparisons to be made concerning instrument format, evidence sequences, and timing of group interaction, as they might affect the assessment curves generated.

The second phase of the study concentrated primarily on the diagnosis process as it occurs in a less highly structured situation. In the second phase, no variations in the conditions of assessment were used. Our analytic interest in this phase was to identify kinds of thinking patterns that were common, and their relation to the social interactions that occurred among the teachers. The present report deals only with the results of the experimental sessions (Phase I).

Developing the Phase I procedures involved several tasks: (1) choosing an experimental design by which to allocate the various treatment combinations; (2) developing the scenario materials; (3) developing the session protocols, and the instruments for measuring outcomes and judgments; and (4) recruiting and scheduling the teacher participants.

The experimental design paradigm that seemed most suitable was a

4 x 4 Latin Square. However, the design was complicated by the fact that each dimension of the square reflected not a single variable having four levels, but instead comprised a 2 x 2 enumeration of two dichotomous variables. The elements of this design are shown in Figure 1.

- - - - - - - - - - -

Figure 1 About Here

- - - - - - - - - - -

The design is a 2^4 arrangement: there are 4 distinct treatment variables under investigation, and each is a dichotomy. Two of these dichotomous variables are applied to entire groups, and the other two are applied to the individual teachers within a group. The third variable in the Latin Square reflects two important background characteristics of a teacher, each of which is also a dichotomy. These are sex and race.

Thus, in Figure 1, each of the 4 columns (separated by double lines) defines the treatment arrangements used for one group of teachers. The first column, headed "Sample Low; Early Interaction," indicates that this group received information indicating relatively low scores from what was described as an actual testing of a sample of students, and held their interaction early in the sequence of activities. So, the first dichotomous variable indicates whether (a) the low evidence input was labeled as being from the testing of a sample, and the higher input was labeled as being from expert judgments made by local teachers, or (b) vice versa. This labeling of the evidence input was the same for all teachers in any group, but was shifted from one group to another. The second variable, also applied to the entire group of teachers, was the

Figure 1

LATIN SQUARE DESIGN

OF PHASE I*

GROUP (VARIABLE 1)

TEACHER (VARIABLE 2)

	Sample Low; Early Interaction ①	Sample Low; Late Interaction ②	Sample High; Early Interaction ③	Sample High; Late Interaction ④
Density Curve; S then E ①	MW	FW	MB	FB
Density Curve; E then S ②	FW	MB	FB	MW
Cumul Curve; S then E ③	MB	FB	MW	FW
Cumul Curve; E then S ④	FB	MW	FW	MB

* Entries in each cell are the levels of SEX-RACE (VARIABLE 3).

timing of the group interaction. For some groups, the interaction was held early in the session, and for other groups it was held later.

The next two dichotomous variables--type of assessment format used (density curve or cumulative curve) and sequence of information (sample-then expert or expert-then sample)--likewise were used to constitute a single 4-level variable for the Latin Square design. The teachers in a group were told explicitly that two different formats were being used for expressing the assessments, but they were not aware that the timing of receipt of the same evidence was different for some than for others.

Finally, because the sex and race of the individual teacher as well as the sex and race composition of the groups might affect the results, these variables were designed into the Latin Square to be constant across all groups and balanced over the within-group factors. Thus, each group contained a white male, a white female, a black male, and a black female. Also, the type of assessment curve used and the sequence of evidence sources experienced were balanced over the sexes and races. A complete cycle of this design thus required 4 groups and 16 teachers.

To understand more clearly the logic of this design, consider the experimental session activities scheduled for group 1. As soon as all four teachers had arrived at the laboratories, they were brought into a conference room and seated in assigned seats around two tables arranged as in Figure 2. The investigator introduced himself, welcomed them to the laboratories and thanked them for participating. Next he outlined the ideas of diagnoses and assessment, and then asked the teachers to fill out Diagnosis Record forms (individually) indicating the clues or

- - - - - - - - - - -

Figure 2 About Here

- - - - - - - - - - -

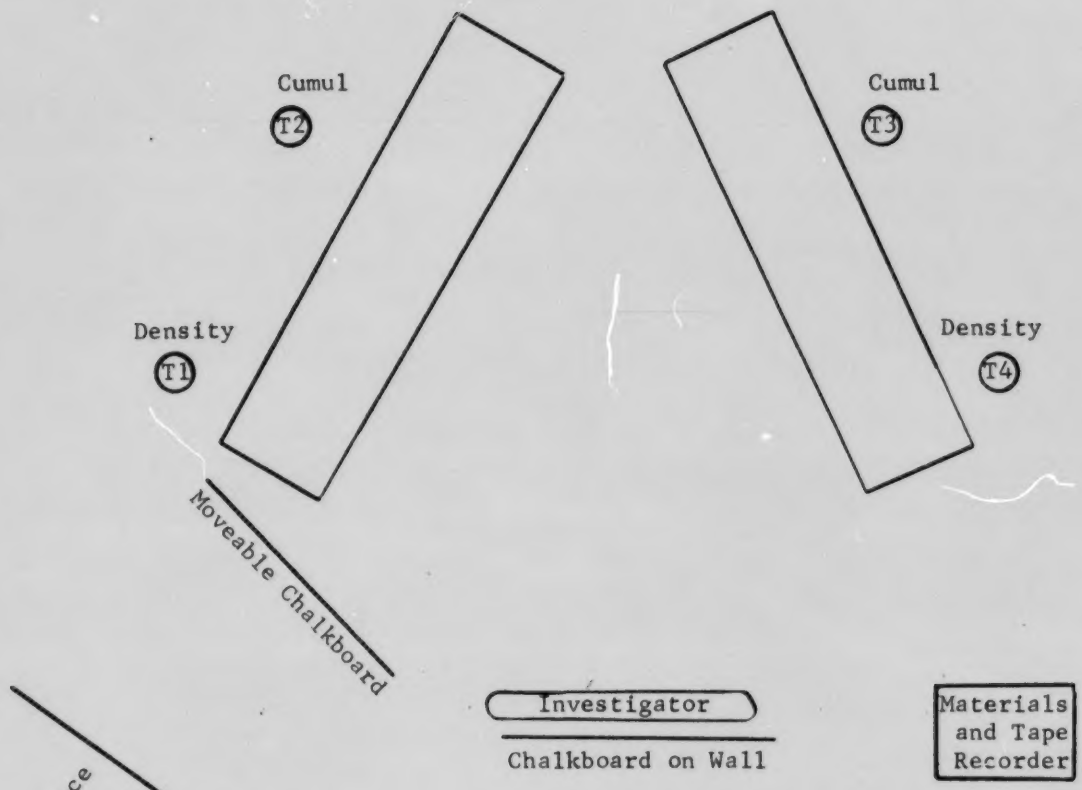
factors they took into account in making judgments about patterns of mathematics performance, and the importance of each such factor. When they had finished that task, each teacher was given a booklet describing a community called Grandview and one of its junior high schools, the Daw School. A copy of each of these booklets is presented in Appendix A. The teachers were asked to read through the booklets and to familiarize themselves with the contents. When they had done this, they were asked to add to their previous answers on the Diagnosis Record Form their judgments as to the level of each relevant factor at the Daw School, based only on the information in the booklets.

At this point, group 1 (and also group 3) were asked to participate in a structured group information sharing. This "early interaction" was held prior to any actual assessments by the teachers, and prior to any receipt by them of direct quantitative evidence about the distribution of mathematics skills among the Daw School students. The interaction was structured to promote nearly uniform participation by each teachers. Each teacher in turn mentioned one "factor" from his/her list, briefly explained it, and then (also briefly) indicated the reasons he/she felt it was relevant.

This structured sharing of information ended the first (diagnosis) part of the session. Our restriction against reactions and evaluative comments by the teachers to each other's ideas in this group sharing was intended to maximize the "available" factors (cf. Tversky and Kahneman,

Figure 2

SEATING LAYOUT



1974) in each teacher's mind.

The next activity of the session was an explanation of the instruments and procedures used to record the assessed distributions, and then a practice assessment (using a quantity unrelated to education). Both the density assessment method and the cumulative method were explained to the entire group. The two teachers seated farthest from the board were assigned to use the cumulative graphs and those nearest the board were assigned to use the density graphs (see Figure 2).

After completing the practice assessment, the teachers had a short rest break, during which they were asked to think about the Daw School 7th graders and what distributions of mathematics skills might be found there.

At the end of the rest break, each teacher was given a Distribution Assessment Form. In each group of four teachers, two teachers received forms designed for the construction of a cumulative distribution curve, and the other two received forms designed for the construction of a density curve. An example of each form is shown in Figure 3. Each form

- - - - - - - - - - -

Figure 3 About Here

- - - - - - - - - - -

showed a grid, on which the horizontal axis indicated levels of mathematics achievement, measured in grade-equivalents. Each teacher was asked to plot the curve which he/she felt best described the distribution of mathematics skills among the 300 seventh-grade students at the Daw School. This task was referred to as the Initial Assessment. In completing it, the teachers used the general information they had ac-

Distribution Assessment Record Form A Initial Assessment identification #_____

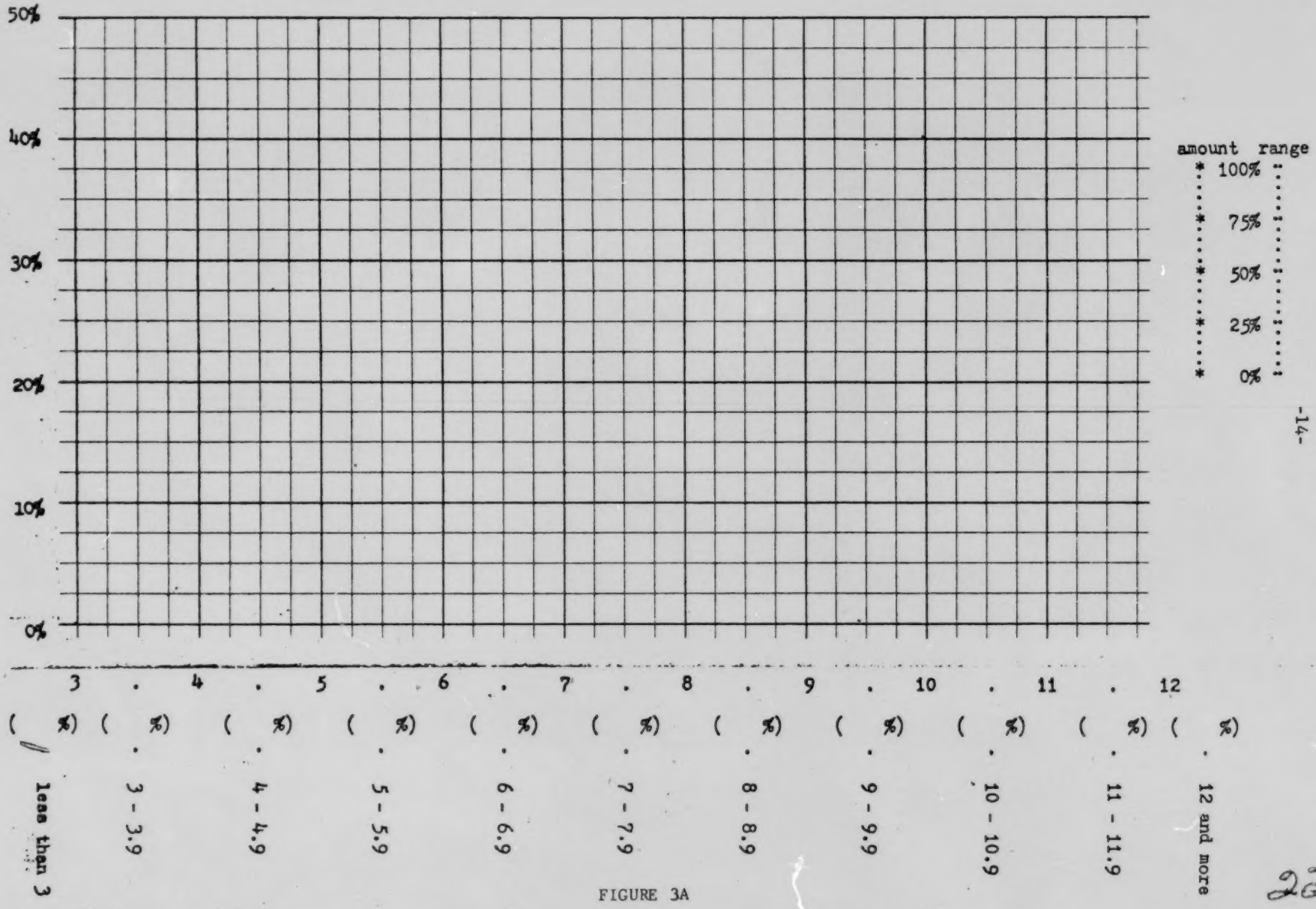


FIGURE 3A

Distribution Assessment Record Form B Initial Assessment identification #_____

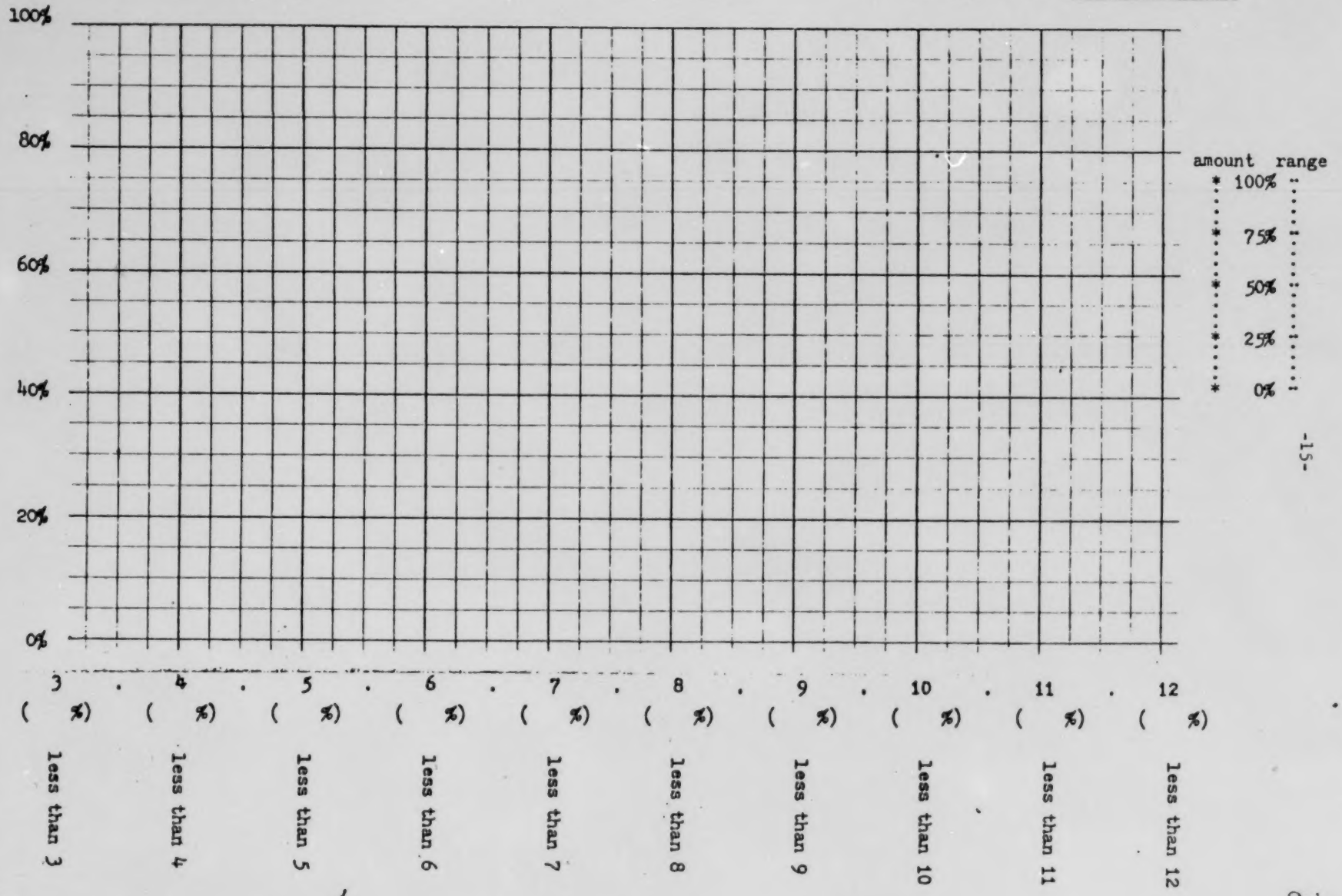


FIGURE 3B

quired from reading the booklets, together with their professional training and experience. They were instructed to assume that this judgment was being made in September, about the entering group of seventh graders.

The two methods of recording a judgment about the distribution of mathematics skills are fully equivalent--a judgment expressed in either method can be easily converted to the other method. The differences in the methods are only in the ease with which human judges (such as the teachers in this study) can make use of them, or in any differential biases the two methods produce in judgments.

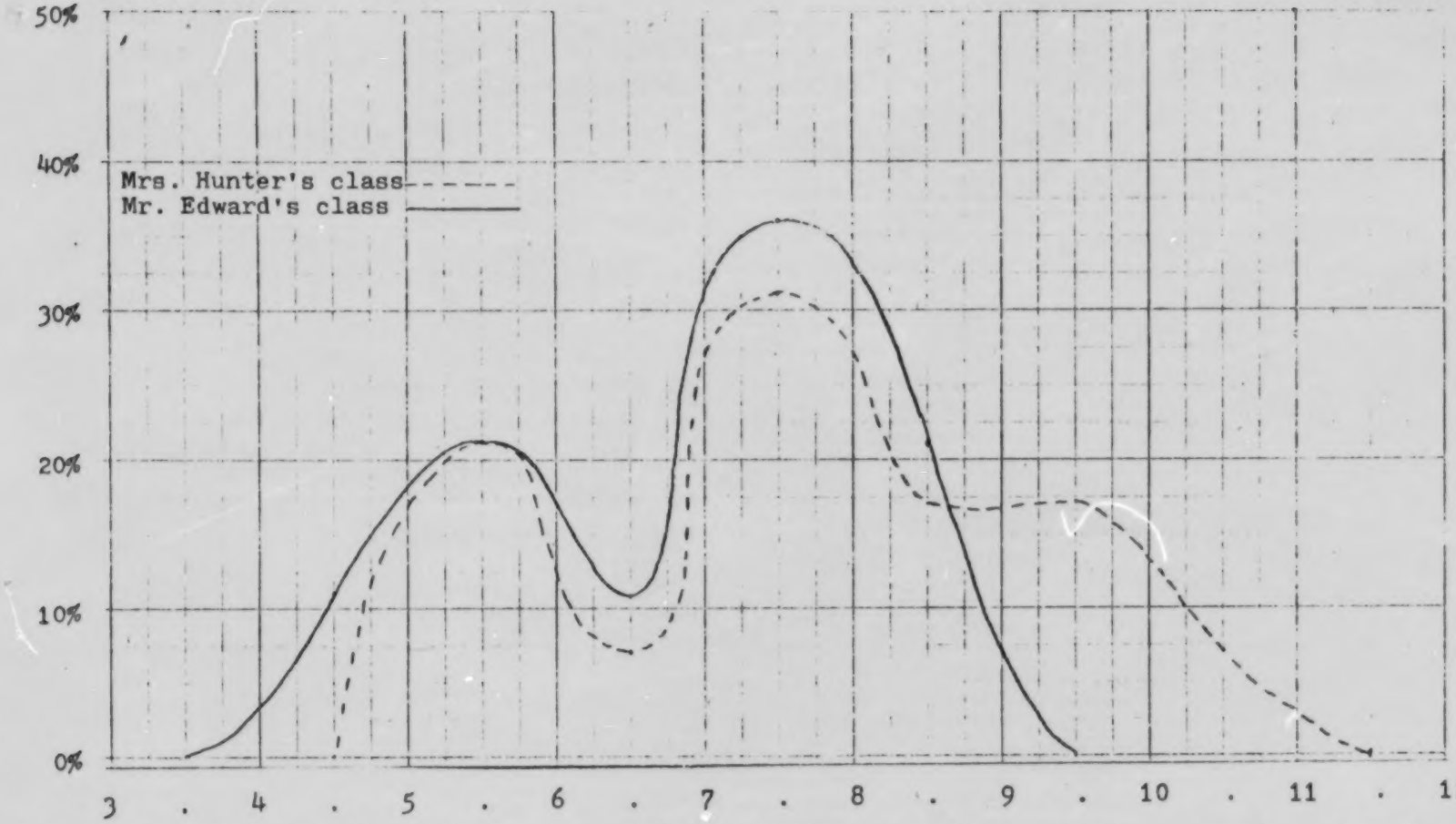
The next part of the session was aimed at ascertaining the reactions by teachers to the receipt of additional information about the mathematics skills of the students. This additional information took the form of distribution curves showing mathematics levels in each of 4 classes of the entire seventh grade (there are 11 classes in the entire grade). The information was presented in two waves, referred to as "Evidence #1 " and "Evidence #2." For "Evidence #1," the teachers were shown two curves on a single sheet; a second set was shown for "Evidence #2." One sheet showed curves from classes taught by Mrs. Hunter and by Mr. Edwards (known as the HE curves). The other sheet showed curves from classes taught by Mrs. Adams and by Mr. Forrester (known as the AF curves). Copies of these curves are shown in Figure 4. The HE curves indicated a lower level of perfor-

- - - - - - - - - - -

Figure 4 About Here

- - - - - - - - - - -

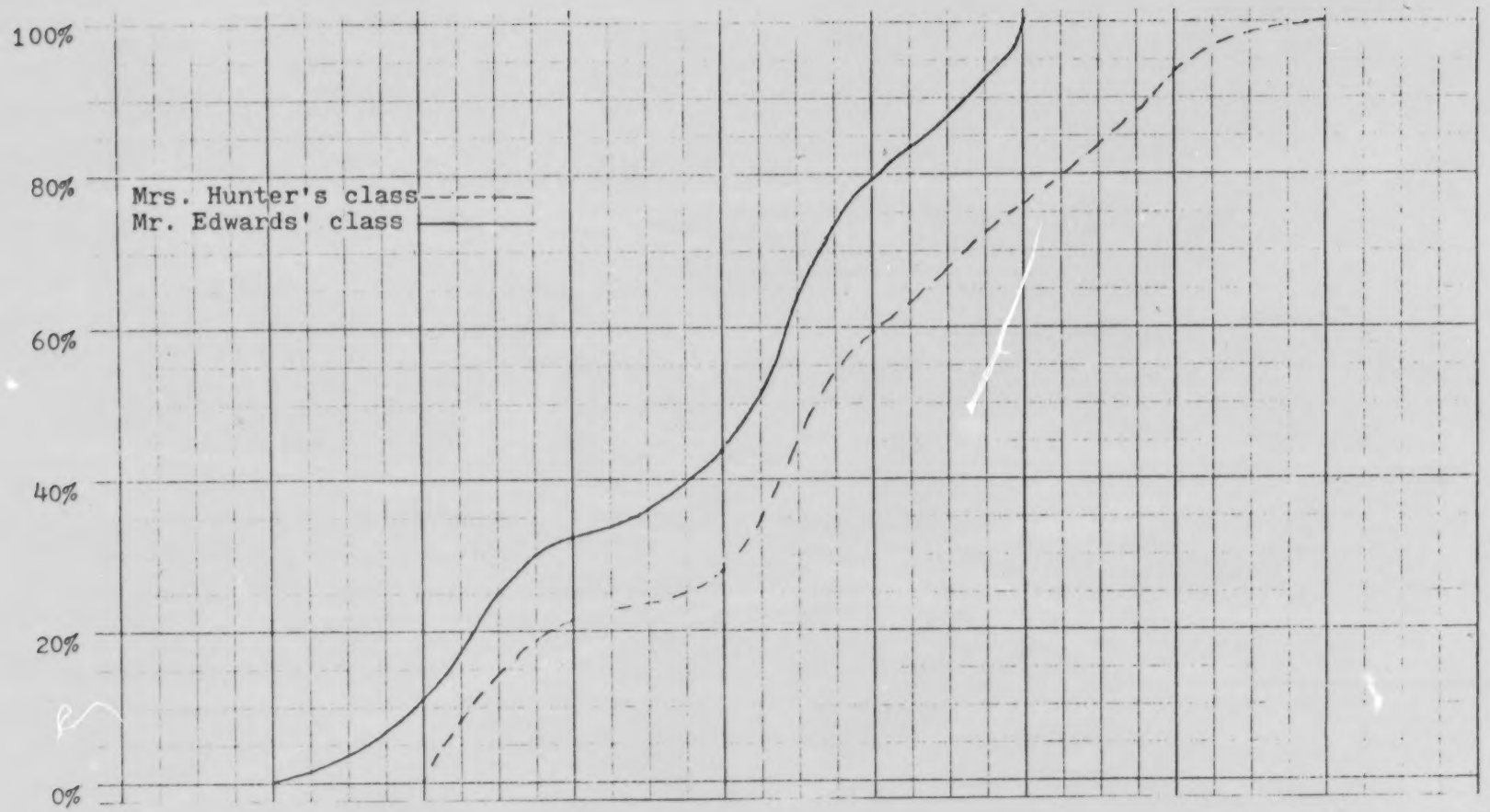
mance than did the AF curves. The delivery sequence and apparent sources of this information were varied across groups and individuals by attaching different descriptive paragraphs to the evidence graphs.



	less than 3 CJ	3 - 3.9	4 - 4.9	5 - 5.9	6 - 6.9	7 - 7.9	8 - 8.9	9 - 9.9	10 - 10.9	11 - 11.9	12 and more CJ
H →	(0 %)	(0 %)	(0 %)	(21 %)	(7 %)	(31 %)	(17 %)	(17 %)	(7 %)	(0 %)	(0 %)
E →	(0 %)	(0 %)	(11 %)	(21 %)	(11 %)	(36 %)	(21 %)	(0 %)	(0 %)	(0 %)	(0 %)

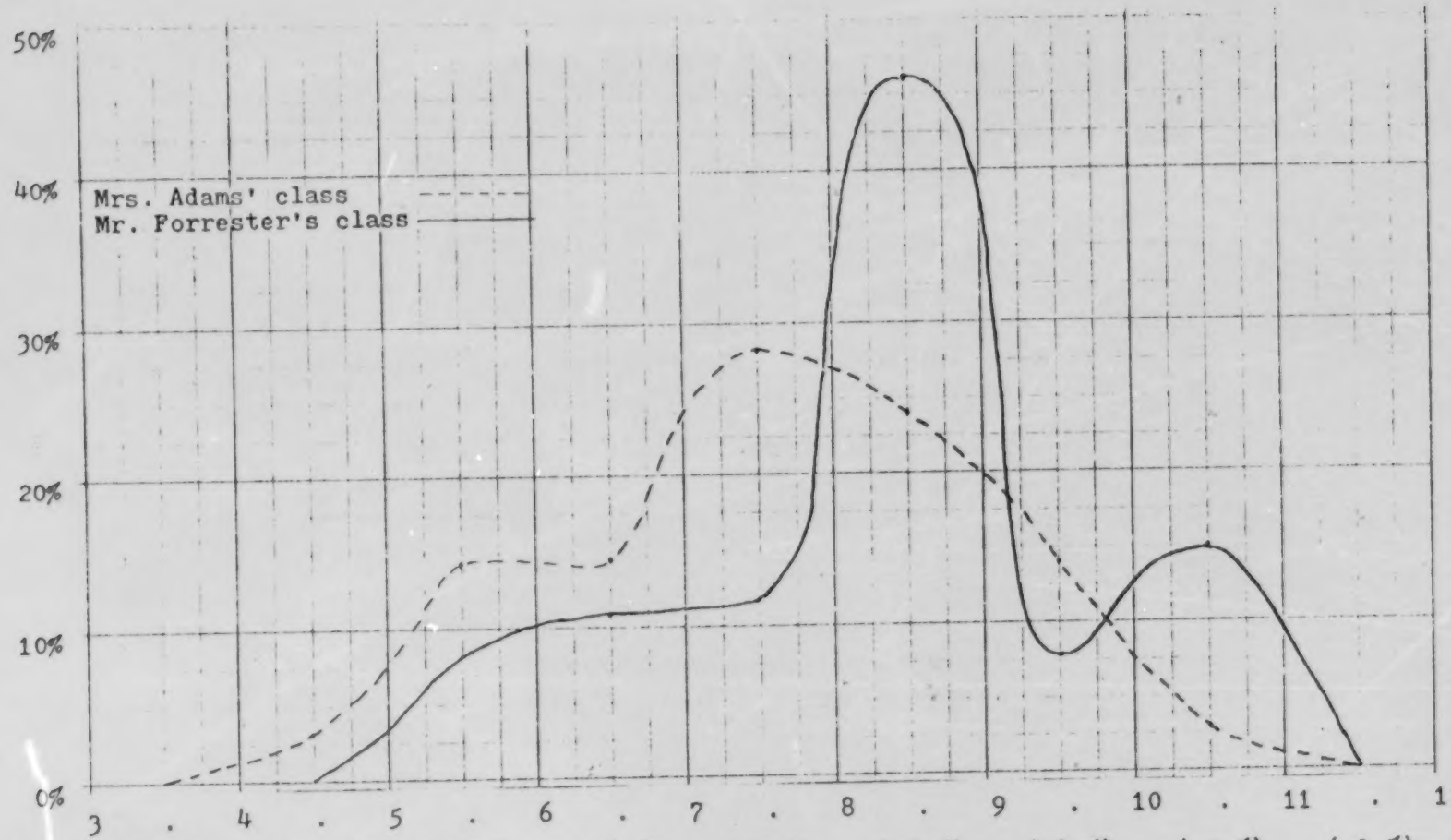
FIGURE 4A

FIGURE 4B



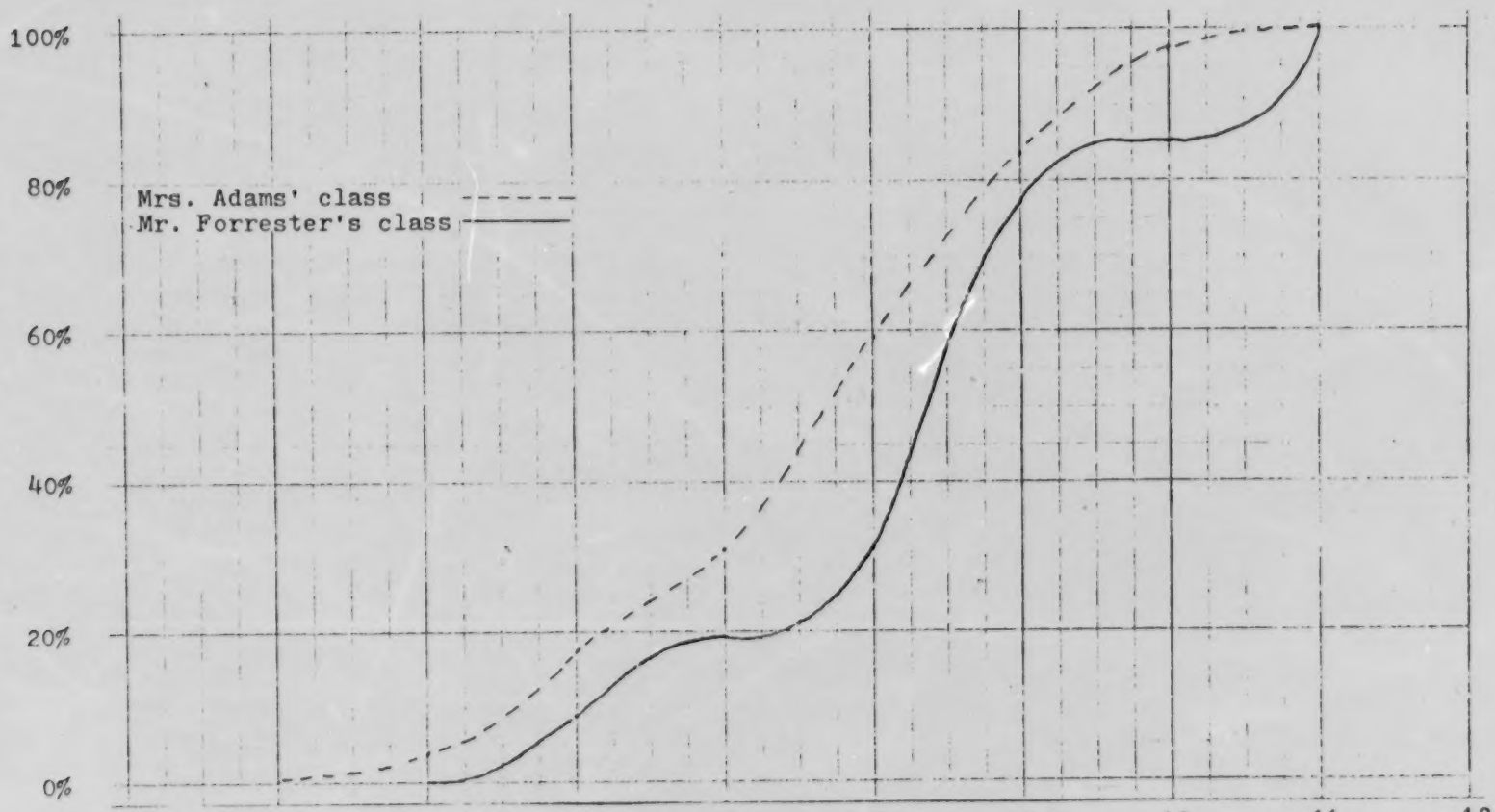
	3	4	5	6	7	8	9	10	11	12
H →	(0 %)	(0 %)	(0 %)	(21%)	(28 %)	(59 %)	(76 %)	(93 %)	(100 %)	(100%)
E →	(0 %)	(0 %)	(11%)	(32%)	(43%)	(79 %)	(100%)	(100%)	(100 %)	(100%)
	less than 3	less than 4	less than 5	less than 6	less than 7	less than 8	less than 9	less than 10	less than 11	less than 12

FIGURE 4C



	less than 3	3 - 3.9	4 - 4.9	5 - 5.9	6 - 6.9	7 - 7.9	8 - 8.9	9 - 9.9	10 - 10.9	11 - 11.9	12 and more
A →	(0 %)	(0 %)	(3 %)	(14 %)	(14 %)	(28 %)	(24 %)	(14 %)	(3 %)	(0 %)	(0 %)
F →	(0 %)	(0 %)	(0 %)	(8 %)	(11 %)	(12 %)	(46 %)	(8 %)	(15 %)	(0 %)	(0 %)

FIGURE 4D


100%
80%
60%
40%
20%
0%
Mrs. Adams' class
Mr. Forrester's class
3 . 4 . 5 . 6 . 7 . 8 . 9 . 10 . 11 . 12
A → (0 %) (0 %) (3 %) (17 %) (31 %) (59 %) (83 %) (97 %) (100 %) (100%)
F → (0 %) (0 %) (0 %) (8 %) (19 %) (31 %) (77 %) (85 %) (100 %) (100%)
less than 3
less than 4
less than 5
less than 6
less than 7
less than 8
less than 9
less than 10
less than 11
less than 12

Thus, in half of the groups the AF graph was labeled as being derived from an actual achievement testing of the (approximately 55) students in those two classes. In the other half of the groups, this label was given to the HE classes data. Conversely, the groups which were told that the AF data came from an achievement testing were also told that the HE data was obtained from the judgments of other teachers like themselves, teachers who had spent some time actually observing a mathematics lesson in each class. This is the "expert" evidence.

Although the four teachers in any single group all had the same source associated with the same data, they did not all receive the same data in the same wave. Thus, the design separates the possible reaction to the perceived source of the information from the possible reaction to the sequence of delivery.

After the first wave of evidence was studied by each teacher, the teachers again filled out individual distribution assessment forms. This was referred to as the Assessment after Evidence #1. The next step was to hand out the second wave of evidence and have it studied. Then another, newly revised assessment was made individually by each of the teachers. This was referred to as the Assessment after Evidence #2. Thus, at this point each teacher individually had made three successive assessments of the distribution of mathematics skills among the Daw School 7th graders. When all the teachers in a group had completed this third judgment (the Assessment after Evidence #2), the forms were collected. The forms were then tabulated by an assistant, using a nearby minicomputer, while the teachers worked for a few minutes writing down explanations of the rela-

tion they saw between the "factors" they had listed as part of their diagnoses and their actual assessments.

As soon as the teachers had completed this task, they were given their last assessment sheet (Assessment after Evidence #2) and a computer printout showing the average distribution assessed by all four of them. These curves were provided in both a density format and a cumulative format. The average curves were explained to the teachers.

At this point, for those groups which had been assigned to the "Late Interaction" treatment, the same procedure of structured group information sharing as described earlier was carried out.

Next the teachers individually were asked to make a fourth and final assessment, based on all the information available to them at that moment. This was referred to as the "Final Assessment." Then, as the last task of the session, they were asked to fill out a short questionnaire indicating some of their reactions to the study and the assessment procedures.

Thus, the available data from Phase I of the project includes the following elements:

1. A Background Questionnaire completed by each teacher.
2. A Diagnosis Record Form completed by each teacher.
3. Four successive Assessment Forms completed by each teacher.
4. A Post-Session Questionnaire completed by each teacher.

The complete cycle of treatments (4 groups of 4 teachers each) was executed for two replications (a total of 32 teachers). Our original intent had been to use more groups, ideally twice as many, but difficulties in matching the groups on sex and race made this impossible.

5. Results of the Phase I Analysis

This section presents an initial analysis of the experimental data, focused on our main research questions. Additional analyses are in progress and will be described in subsequent reports.

Although our experimental design followed a conventional Latin Square paradigm, the assessment outcome measures are not directly analyzable by univariate analysis of variance techniques because a distribution curve is not a single numerical quantity. Also, the definition of the treatment factors used in the Latin Square is somewhat complex, making discussion cumbersome. Because of these features of the study, we have chosen a more intuitive framework of analysis.

We will present separate comparisons of each treatment factor of interest, rather than a simultaneous analysis of all the factors. This approach sacrifices some of the potential precision of the Latin Square design, but still capitalizes upon the control provided by that design over possible confounding factors. In other words, because we have used a Latin Square design, we can be sure that our discussion of the impact of any single treatment factor is not distorted by the possible main effects of any of the other factors. Each treatment factor is balanced across all of the others. However, by calculating the separate comparisons, we will obtain somewhat larger estimates of disturbance variance than is absolutely necessary, and so may not detect some small effects.

The assessed distributions (the outcome variable of the experiment) are represented in two ways. The simplest and most direct way to summarize a smooth distribution curve is by a measure of its central tendency and a measure of its dispersion. Thus, as one way to analyze the distribu-

tions, we computed the mean and standard deviation for each distribution across the treatment levels of interest. For Normal-distribution curves, such as these, the two statistics are statistically independent. Thus, the separate analyses of two outcomes is, for this particular set of measures, approximately treatable as two univariate outcomes.

The assessment curve means were averaged over all the teachers in the same set of conditions. Similarly, for each teacher in the same set of conditions, the assessment curve standard deviations were averaged. Thus we are comparing the mean of distribution means across teachers, for two sets of teachers. Each of the means under comparison has its own standard deviation, and this standard deviation is used in judging the magnitude of the difference in the assessment curve means between the two sets of teachers. Similarly, we are comparing the mean of the assessment curve standard deviations across teachers, for two sets of teachers. This mean standard deviation has its own standard deviation, which is used in judging the magnitude of these comparisons.

A second way to summarize the distribution provided by each teacher is simply to report, for each teacher, a set of eleven numbers. Each number indicates the percentage of cases assigned by that teacher to each of the eleven categories of mathematics skill. Because these eleven numbers must add to 100 percent, only 10 of them are analytically independent outcomes. Also, it should be clear that these numbers are <u>not</u> statistically independent. Thus, the strictly formal analysis of such data should be a 10-outcome multivariate comparison. Again, however, we felt that it was more important to present the patterns in some detail, and to indicate where the differences, if any, do appear, than to

obtain precise estimates of the statistical significance of a given comparison. In other words, our aim is more to describe than to test. For each comparison to be discussed, we have presented not only the assessment curve means and standard deviations, but also the mean percentages (as well as their standard deviations and a univariate F statistic useful as a summary descriptive index), we have also provided graphs showing the comparative distributions. These pictorial summary presentations, though least precise, are perhaps most informative as to the actual impact, if any, of the different treatments.

Our analysis in this report is confined to three of the treatment variables. First, we examined the impact of using the density format to make the assessment as contrasted with use of the cumulative format. Second, we considered the assessed distributions among teachers who experienced the structured sharing of information early, as compared with the same distributions among the teachers who experienced this social interaction at a later stage in the session. Third, we compared the reactions by teachers to the perceived source of a particular item of information. That is, are the assessed curves different when information is labeled as coming from an objective testing in contrast to being labeled as the judgment of other teachers? We shall also discuss a pattern of aggregate accuracy and individual idiosyncrasy that was observed as an overall feature of the analysis.

5.1 The Impact of Different Assessment Methods

As described earlier, half the teachers in the study used a "density" method to assess the distribution of mathematics skills, and the other half used a "cumulative" method. To present comparative results,

we have expressed all distributions in the density form, but this is a strictly mathematical transformation. Our interest in this comparison is in whether the distributions obtained from the teachers are affected by the kind of instrument format they were using. The results of our analysis are given first in terms of the assessment curve means and standard deviations under each format, and then in terms of the actual percentages assigned to each category under each format.

Table 1 gives the means and standard deviations of the assessed distributions, at each of the four times, for the 16 teachers who used the density method and for the 16 teachers who used the cumulative method. Also, for each comparison, an F-statistic is presented which would be

- - - - - - - - - - -

Table 1 About Here

- - - - - - - - - - -

the appropriate statistic for testing a hypothesis of no difference in the quantities under consideration, if the necessary statistical assumptions were satisfied. Strictly speaking, such assumptions are not satisfied, because the means are repeated measures on the same persons and thus not statistically independent. Also, there is no assurance that the observed quantities are drawn from a Normal distribution. Nevertheless, the F-statistic does provide a convenient descriptive summary index of the magnitude of the difference in relation to the standard deviation of each score, and is presented for that purpose.

Table 1 has two panels. In Panel A, the assessment curve means for all teachers who used the density method, and for all teachers who used the cumulative method, are averaged for each of the four time points.

Table 1

Density Method vs. Cumulative Method
Comparison of Means and Sigmas

A. Means of the Assessed Curves

treatment	Initial	After info #1	After info #2	Final	
density	6.749	7.746	7.377	7.370	N = 16
cumulative	6.465	7.236	7.378	7.322	N = 16
F-statistic	0.718	2.791	0.000	0.241	

B. Sigmas of the Assessed Curves

treatment	Initial	After info #1	After info #2	Final	
density	1.674	1.602	1.560	1.610	N = 16
cumulative	2.110	1.733	1.613	1.666	N = 16
F-statistic	5.805	2.634	0.593	0.492	

Thus, among the 16 teachers who used the density method to represent their judgments, the mean of their individual distribution means was 6.749 at the initial assessment and 7.370 at the final assessment. Among the 16 teachers who used the cumulative method, the mean at the initial assessment was 6.465 and 7.322 at the final assessment. The F-statistic for the comparisons at the initial and final assessments are quite small, 0.718 and 0.241 respectively. These F-statistics provide a handy summary index of the size of the differences in the two groups.

Panel B of Table 1 shows the corresponding comparisons for the standard deviations of the distributions provided by the two groups of teachers. Among the 16 teachers who used the density method, at the initial assessment the average of the 16 sigmas was 1.674. Among the other 16 teachers, the corresponding average was 2.110. The F-statistic for this particular comparison is 5.805, suggesting that the difference in sigmas, at this initial assessment, is sizeable.

The results given in Table 1 show that there may be some tendency for the cumulative method to produce a wider distribution, but that this tendency decreases as more information is acquired. The means of the assessment curves, at each time point, are quite similar for the two groups of teachers, although the teachers using the cumulative method have a lower mean at three of the four times. The cumulative method teachers also have a greater assessment curve sigma at each of the four times.

A more detailed comparison of the differences between the two methods is given in Table 2, and in Figures 5, 6, 7, and 8. Table 2 shows the

- - - - - - - - - - - - - - - - - -

Table 2 About Here
Figures 5, 6, 7, and 8 About Here

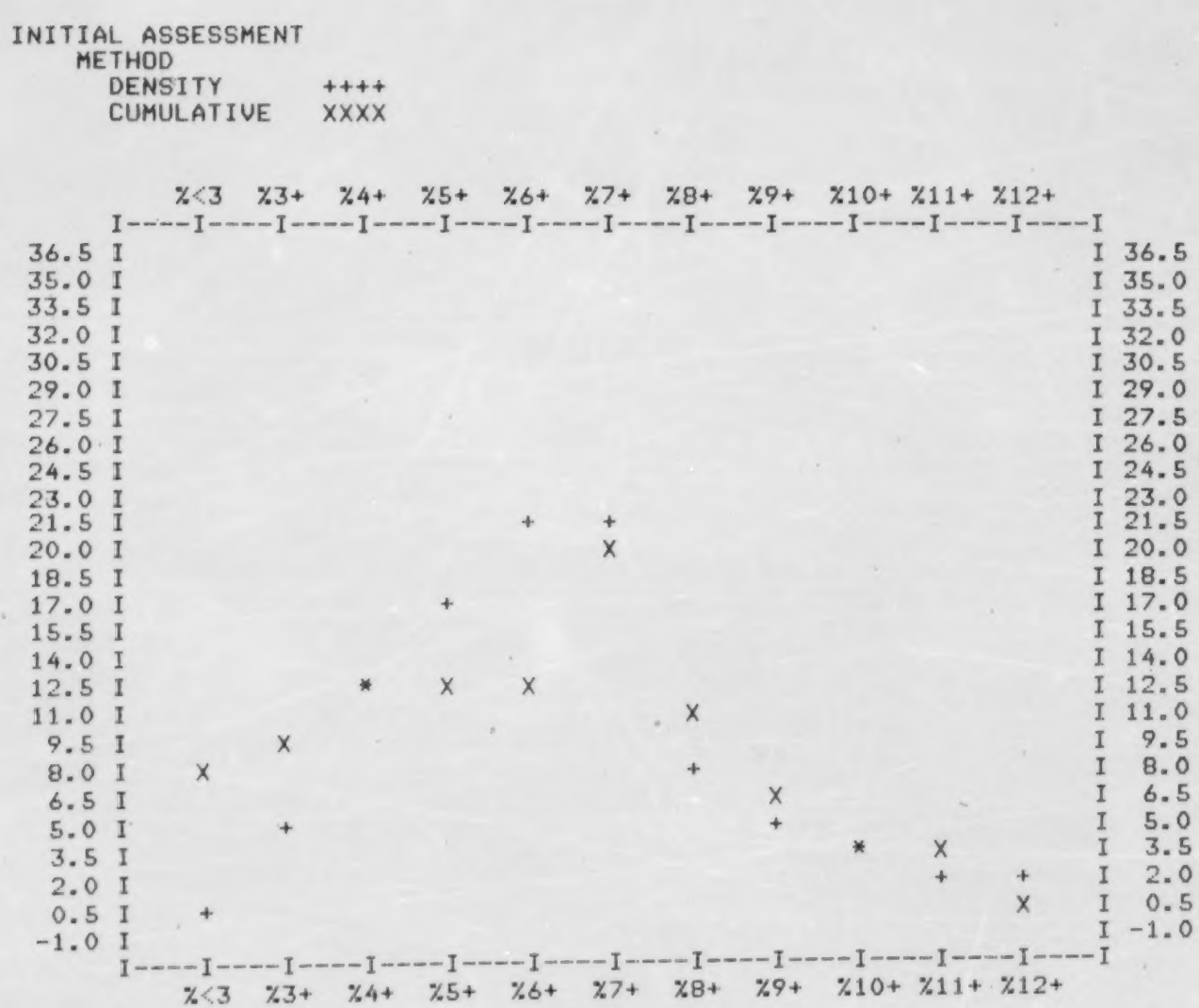
- - - - - - - - - - - - - - - - - -

Table 2

Density Method vs. Cumulative Method
Comparison of Percentage Assignments

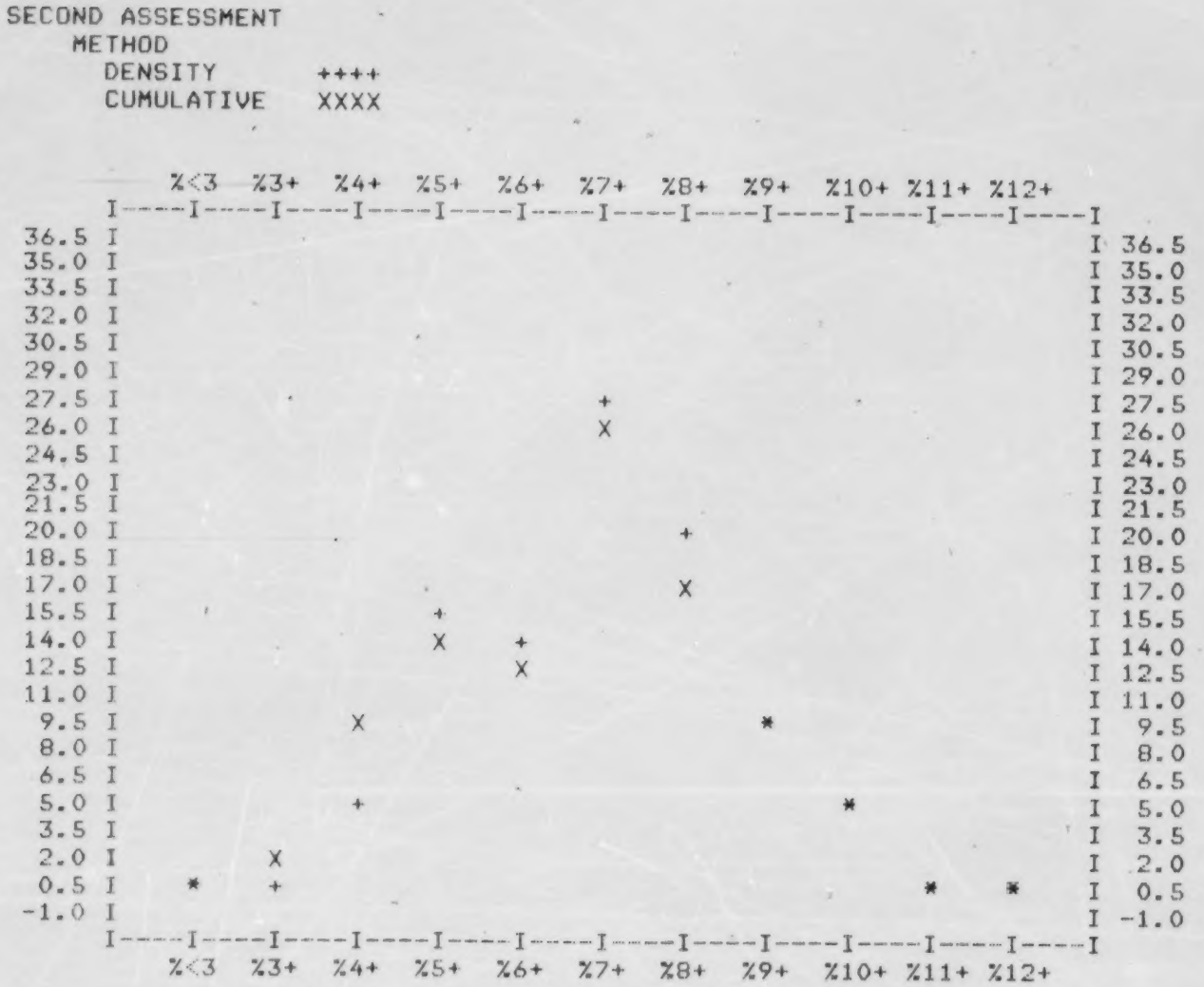
	3.0 -	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-9.9	10-10.9	11-11.9	12 +
Initial Assessment											
density method	1.219	5.717	12.025	16.804	21.671	21.440	8.659	5.512	3.609	2.000	1.344
cumulative method	7.876	9.669	11.894	13.063	12.563	20.625	10.688	6.438	4.719	3.156	0.313
F-statistic	10.154	2.797	0.002	2.279	10.040	0.023	1.086	0.288	0.004	0.383	1.217
After info #1											
density method	0.131	0.886	5.345	14.864	14.098	27.643	20.528	10.133	5.139	0.754	0.478
cumulative method	0.875	1.875	9.875	14.375	12.938	26.250	17.563	10.125	5.531	0.594	0.000
F-statistic	3.741	1.910	6.545	0.081	0.284	0.228	0.776	0.000	0.082	0.114	1.881
After info #2											
density method	0.006	0.781	5.525	16.656	13.813	29.688	18.938	9.281	4.525	0.469	0.319
cumulative method	0.250	1.000	6.938	15.188	12.625	28.500	20.375	9.250	5.531	0.344	0.000
F-statistic	1.583	0.100	1.650	1.191	0.268	0.426	0.272	0.001	0.797	0.107	1.043
Final Assessment											
density method	0.069	1.094	6.189	15.966	14.429	27.301	19.765	9.701	4.701	0.531	0.256
cumulative method	0.344	1.375	9.375	13.000	12.938	27.563	20.125	9.938	4.875	0.469	0.000
F-statistic	1.476	0.186	3.543	3.970	0.461	0.007	0.025	0.043	0.030	0.026	1.759

FIGURE 5



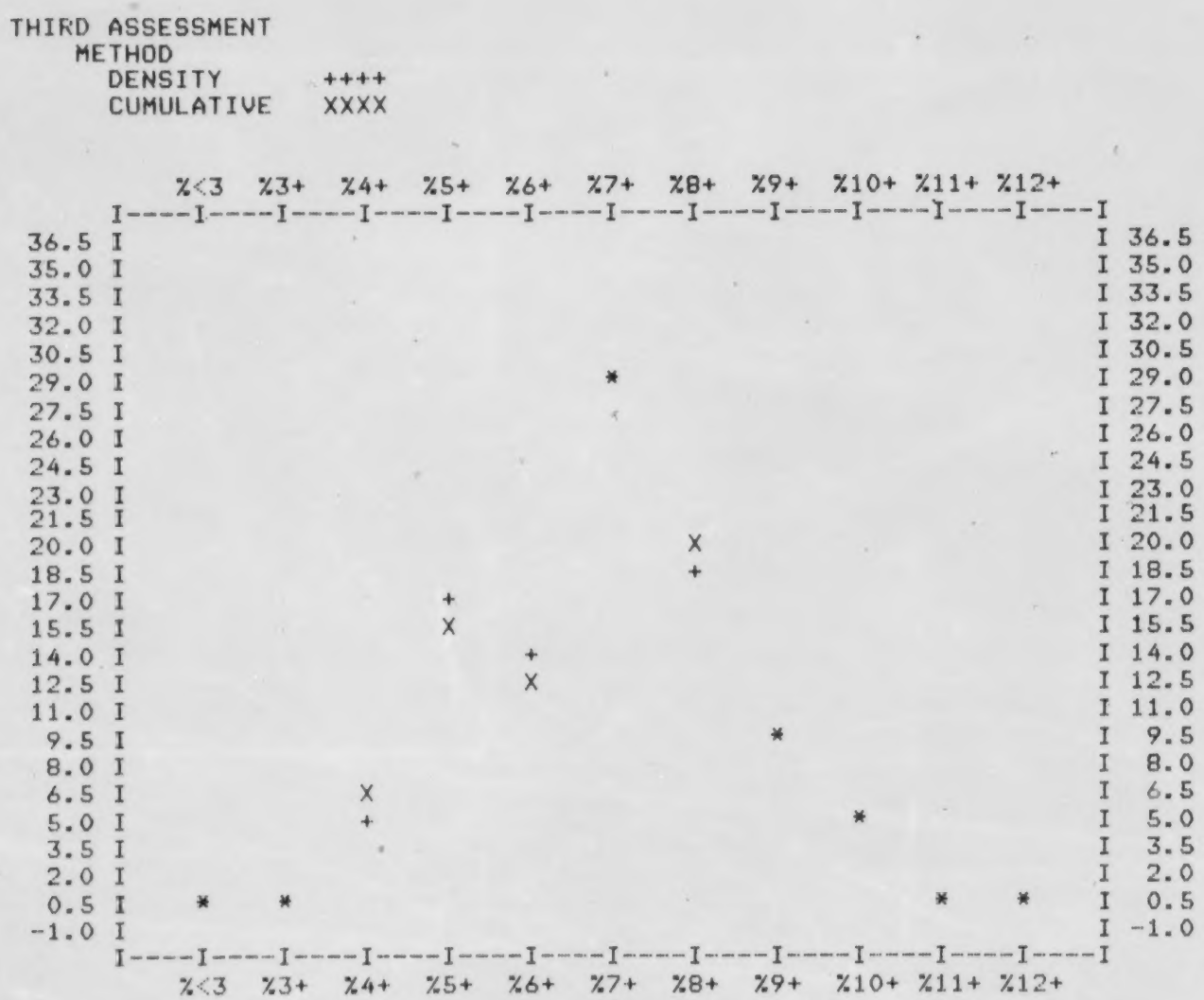
21.671	5	20.625	6	*	1	1.219	7.875
21.440	6	13.063	4	*	2	5.717	9.669
16.804	4	12.563	5	*	3	12.024	11.894
12.024	3	11.894	3	*	4	16.804	13.063
8.659	7	10.688	7	*	5	21.671	12.563
5.717	2	9.669	2	*	6	21.440	20.625
5.512	8	7.875	1	*	7	8.659	10.688
3.609	9	6.438	8	*	8	5.512	6.438
2.000	10	3.719	9	*	9	3.609	3.719
1.344	11	3.156	10	*	10	2.000	3.156
1.219	1	0.313	11	*	11	1.344	0.313

FIGURE 6



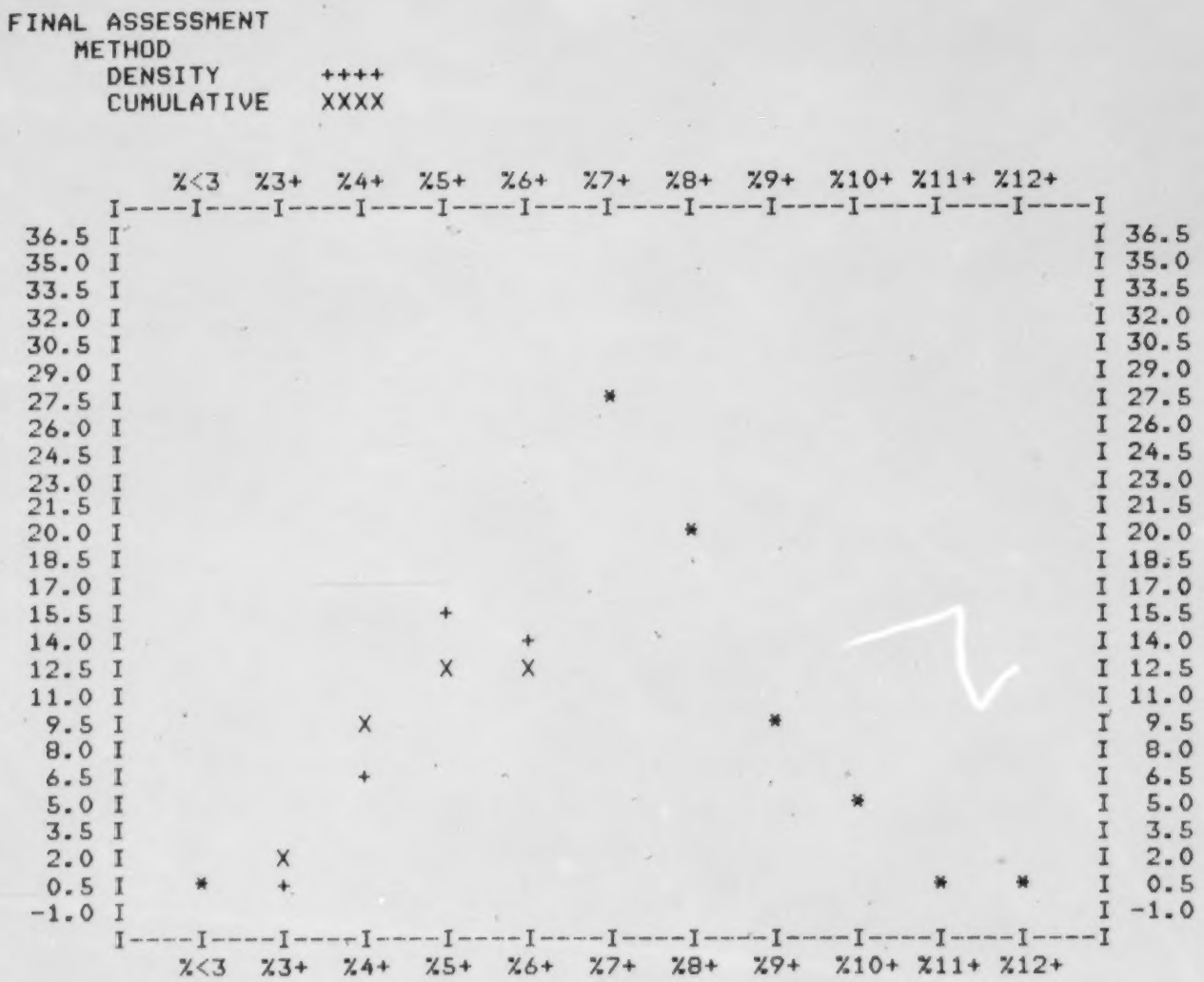
27.643	6		26.250	6	*	1	0.131	0.875
20.528	7		17.563	7	*	2	0.886	1.875
14.864	4		14.375	4	*	3	5.345	9.875
14.098	5		12.938	5	*	4	14.864	14.375
10.133	8		10.125	8	*	5	14.098	12.938
5.345	3		9.875	3	*	6	27.643	26.250
5.139	9		5.531	9	*	7	20.528	17.563
0.886	2		1.875	2	*	8	10.133	10.125
0.753	10		0.875	1	*	9	5.139	5.531
0.478	11		0.594	10	*	10	0.753	0.594
0.131	1		0.000	11	*	11	0.478	0.000

FIGURE 7



29.688	6	28.500	6	*	1	0.006	0.250
18.938	7	20.375	7	*	2	0.781	1.000
16.656	4	15.188	4	*	3	5.525	6.938
13.813	5	12.625	5	*	4	16.656	15.188
9.281	8	9.250	8	*	5	13.813	12.625
5.525	3	6.938	3	*	6	29.688	28.500
4.525	9	5.531	9	*	7	18.938	20.375
0.781	2	1.000	2	*	8	9.281	9.250
0.469	10	0.344	10	*	9	4.525	5.531
0.319	11	0.250	1	*	10	0.469	0.344
0.006	1	0.000	11	*	11	0.319	0.000

FIGURE 8



27.301	6	27.563	6	*	1	0.069	0.344
19.765	7	20.125	7	*	2	1.094	1.375
15.966	4	13.000	4	*	3	6.189	9.375
14.429	5	12.938	5	*	4	15.966	13.000
9.701	8	9.938	8	*	5	14.429	12.938
6.189	3	9.375	3	*	6	27.301	27.563
4.701	9	4.875	9	*	7	19.765	20.125
1.094	2	1.375	2	*	8	9.701	9.938
0.531	10	0.469	10	*	9	4.701	4.875
0.256	11	0.344	1	*	10	0.531	0.469
0.069	1	0.000	11	*	11	0.256	0.000

average percentage judged to be in each category of mathematics skills by the two groups of teachers, at each of the four times. These mean percentages, and those in later tables, are presented in terms of the density method. That is, Table 2 indicates that, among the teachers using the density method, at time 1, the mean of the percentages assigned to the grade-equivalent category "less than 3.0" is 1.219 percent. For the teachers using the cumulative method, the corresponding mean percentage is 7.875 percent. Similarly, for the Final Assessment, the 16 teachers using the density method of assessment assigned 14.429 percent of the cases to the score range 6.0 through 6.9; the other 16 teachers, using the cumulative assessment method, assigned 12.938 percent of the students to the same score interval.

The table shows a tendency, especially at the initial assessment, for the teachers using the cumulative method to provide a wider or flatter and slightly lower assessment. The graph shown in Figure 5 confirms this impression. Thus, for the score categories 4.0 through 7.9, the teachers using the density method assigned more cases, but for the categories 3.0 through 3.9 and 8.0 through 11.9 the reverse is true.

These differences in distribution nearly disappear in the second assessment, which was made after receipt of the first wave of quantitative information. This suggests that the experience of seeing an actual curve helps to anchor the teachers' judgments considerably, even though the curve presented gives only partial information about the actual situation. Similarly, the assessment curves after information #2 and those of the Final Assessment are very close for the two groups of teachers. This pro-

vides further evidence that the distorting potential of the format occurs only in the absence of any close exposure to a similar set of results. We can conclude that the cumulative method tends to encourage an excess dispersion and to encourage placement of an excess proportion of cases in the leftmost parts of the assessed distribution. This tendency is detectable, however, only in the initial assessment. The corresponding graphs shown in Figures 6, 7, and 8 confirm these conclusions.

5.2 The Impact of Early vs. Late Information Sharing

Various arguments can be made about the likely consequences of having the structured information sharing occur early (before any of the assessments are actually made), or late (after all the quantitative information has been received, and several assessments have been made). It may be that early sharing will promote consensus among the teachers and a narrow average distribution, as the teachers tend to concentrate their attention on the ideas raised in the discussion, which will tend to describe typical children rather than unusual children. Conversely, it may be that the introduction of additional ideas at an early stage will promote diversity of opinion and a wide average distribution, as each teacher now bases his/her assessment on a greater variety of facts. Both lines of reasoning are plausible, but they point in opposite directions.

To investigate empirically the consequences of having the structured information sharing early versus late, we carried out a series of comparisons similar in form to the ones used for the comparison of the two instrument formats. In the comparison of formats, the group context was automatically held constant, because the same mix of methods was used

within each group. For analysis of the timing of information sharing, however, there is a possible confounding with the unique group characteristics of each of the eight groups, because the group interaction timing was applied uniformly to all the members of any particular group.

Table 3 shows the comparison of assessment curve means and sigmas for the average distribution assessed by teachers in each information sharing condition, at each of the four assessments. For the first three

- - - - - - - - - -

Table 3 About Here

- - - - - - - - - -

assessments, the means are quite similar. At the final assessment, however, the difference between the two groups is quite large. More important than this magnitude, perhaps, is the fact that the trend among the late sharing group has been upward until time 4, and then it turns downward, even though the trend among the early sharing group remains upward. This seems to be fairly clear evidence that the late discussion (which was held between assessment 3 and assessment 4 for the late group) caused the members of those groups to lower their mean assessment somewhat. Although the interpretation is not fully clear, it may be that early sharing caused the teachers in the early sharing groups to lower their judgments slightly at the first assessment. Thus, the group interaction seems to reinforce the tendency to rely on one's prior opinions, and to attach less weight to the direct evidence.

With respect to the dispersion of the assessed curves among the two sets of teachers, the picture is fairly clear. The teachers in the late sharing groups consistently show wider distributions. This may indicate

Table 3

Early Information Sharing vs. Late Information Sharing
Comparison of Means and Sigmas

A. Means of the Assessed Curves

Treatment	Initial	After info #1	After info #2	Final
early sharing	6.583	7.407	7.415	7.447
late sharing	6.631	7.304	7.339	7.246
F-statistic	0.020	0.476	0.517	4.816

B. Sigmas of the Assessed Curves

Treatment	Initial	After info #1	After info #2	Final
early sharing	1.741	1.581	1.544	1.559
late sharing	2.044	1.754	1.629	1.717
F-statistic	2.559	4.915	1.577	4.352

that the early sharing tends to narrow the range of scores considered likely by the teachers, but other interpretations are also possible. The fact that the assessment curve sigma among the late sharing teachers increases from assessment 3 to assessment 4 seems to be contrary to the interpretation just made, and indicates perhaps that the sharing tended (at least for this group) to increase the width of the distribution. However, the data on this are weak. Our best interpretation is that the interaction tended to narrow the range somewhat, and that the reversal just described is a statistical fluctuation.

We also compared the actual distributions assessed by the teachers in each timing condition. The results of these four comparisons are presented in Table 4, and in Figures 9, 10, 11, and 12.

- - - - - - - - - - - - - - - - - - -

Table 4 About Here
Figures 9, 10, 11, and 12 About Here

- - - - - - - - - - - - - - - - - - -

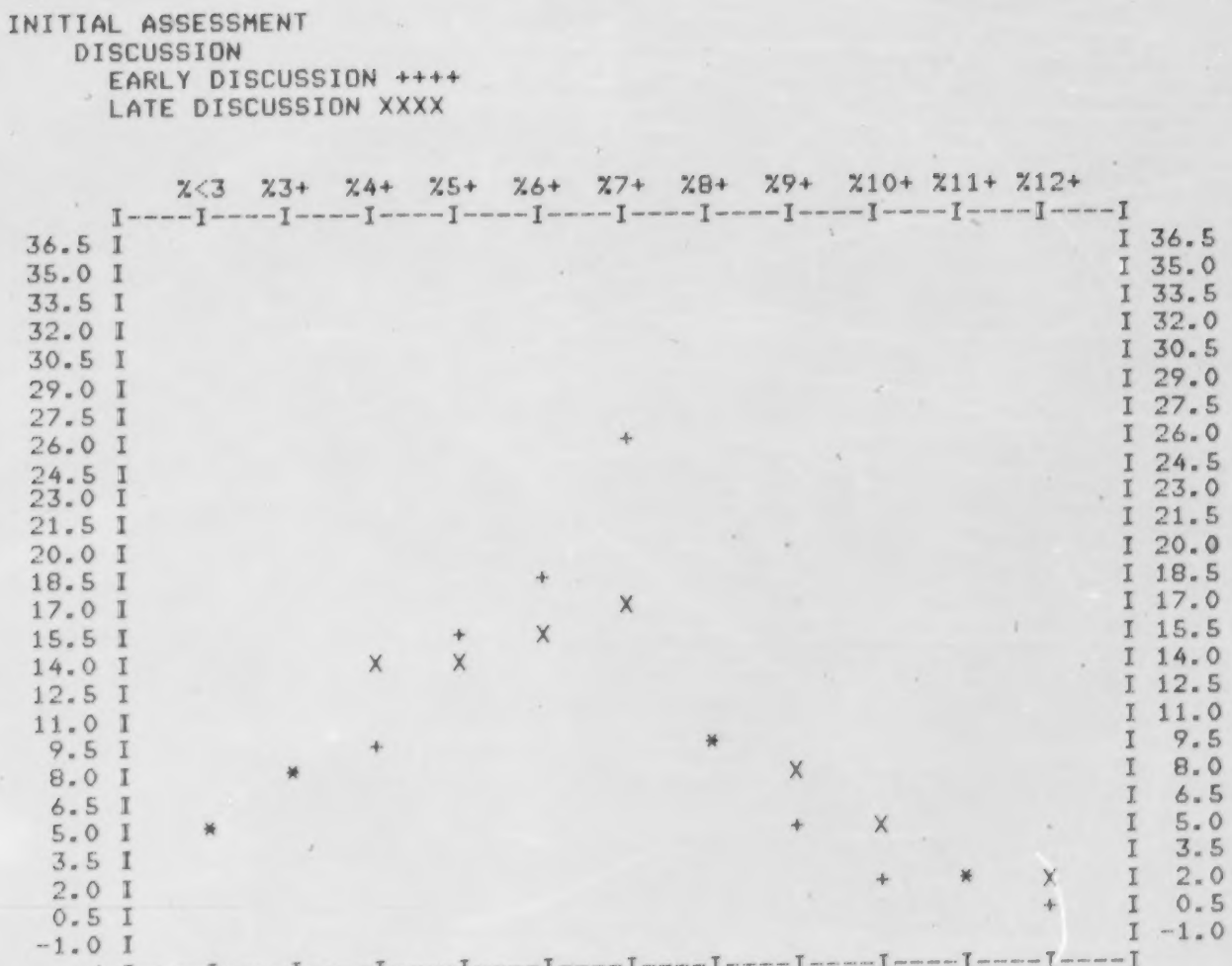
The results of the four comparisons confirm the indications given in Table 3. For the first three comparisons, the shapes are quite similar, but the teachers in the late sharing condition have slightly wider dispersion and a slightly lower location along the grade-equivalent range (i.e., a greater percentage of cases in the lower categories, and a lesser percentage of cases in the higher categories). Figure 12 indicates that at the final assessment, the teachers in the late sharing condition shifted to place a greater percentage of cases in the range 3.0 through 6.9. This is consistent with the finding from Table 3 that the assessment curve mean among the late sharing teachers moves down between time 3 and time 4, and that the standard deviation of the average curve increases. These

Table 4

Early Sharing vs. Late Sharing
Comparison of Percentage Assignments

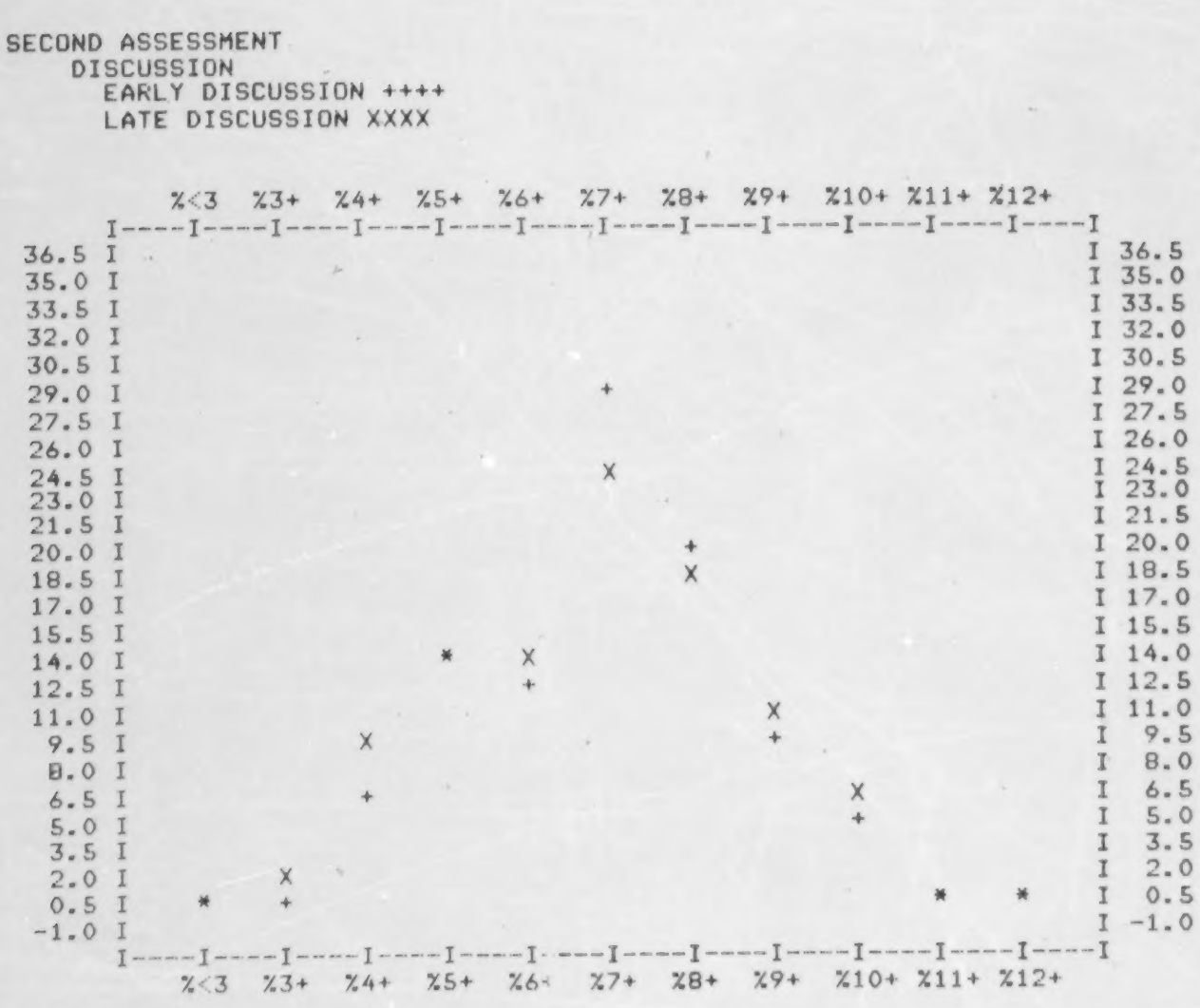
Initial Assessment	3.0 -	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-9.9	10-10.9	11-11.9	12 +
early sharing	4.281	7.356	9.581	15.375	18.500	25.750	9.188	4.688	2.625	2.563	0.094
late sharing	4.813	8.030	14.337	14.492	15.734	16.315	10.159	7.262	4.702	2.594	1.563
F-statistic	0.048	0.075	2.264	0.119	0.710	3.504	0.242	2.387	1.454	0.000	2.575
After info #1											
early sharing	0.194	1.094	6.216	14.603	13.138	29.515	19.766	9.911	4.870	0.625	0.069
late sharing	0.813	1.667	9.005	14.637	13.898	24.378	18.324	10.346	5.800	0.722	0.410
F-statistic	2.494	0.614	2.185	0.000	0.121	3.422	0.180	0.098	0.465	0.042	0.927
After info #2											
early sharing	0.069	0.469	5.963	15.469	12.313	30.563	20.438	9.969	4.494	0.250	0.006
late sharing	0.188	0.313	6.500	16.375	14.125	27.625	18.875	8.563	5.563	0.563	0.313
F-statistic	0.361	1.565	0.228	0.443	0.631	2.807	0.321	1.174	0.902	0.681	0.960
Final Assessment											
early sharing	0.225	1.094	5.275	13.719	13.125	30.938	20.188	10.406	4.650	0.313	0.069
late sharing	0.188	1.375	10.289	15.247	14.241	23.926	19.702	9.232	4.926	0.688	0.188
F-statistic	0.026	0.186	10.623	0.961	0.257	5.706	0.045	1.105	0.074	0.982	0.361

FIGURE 9



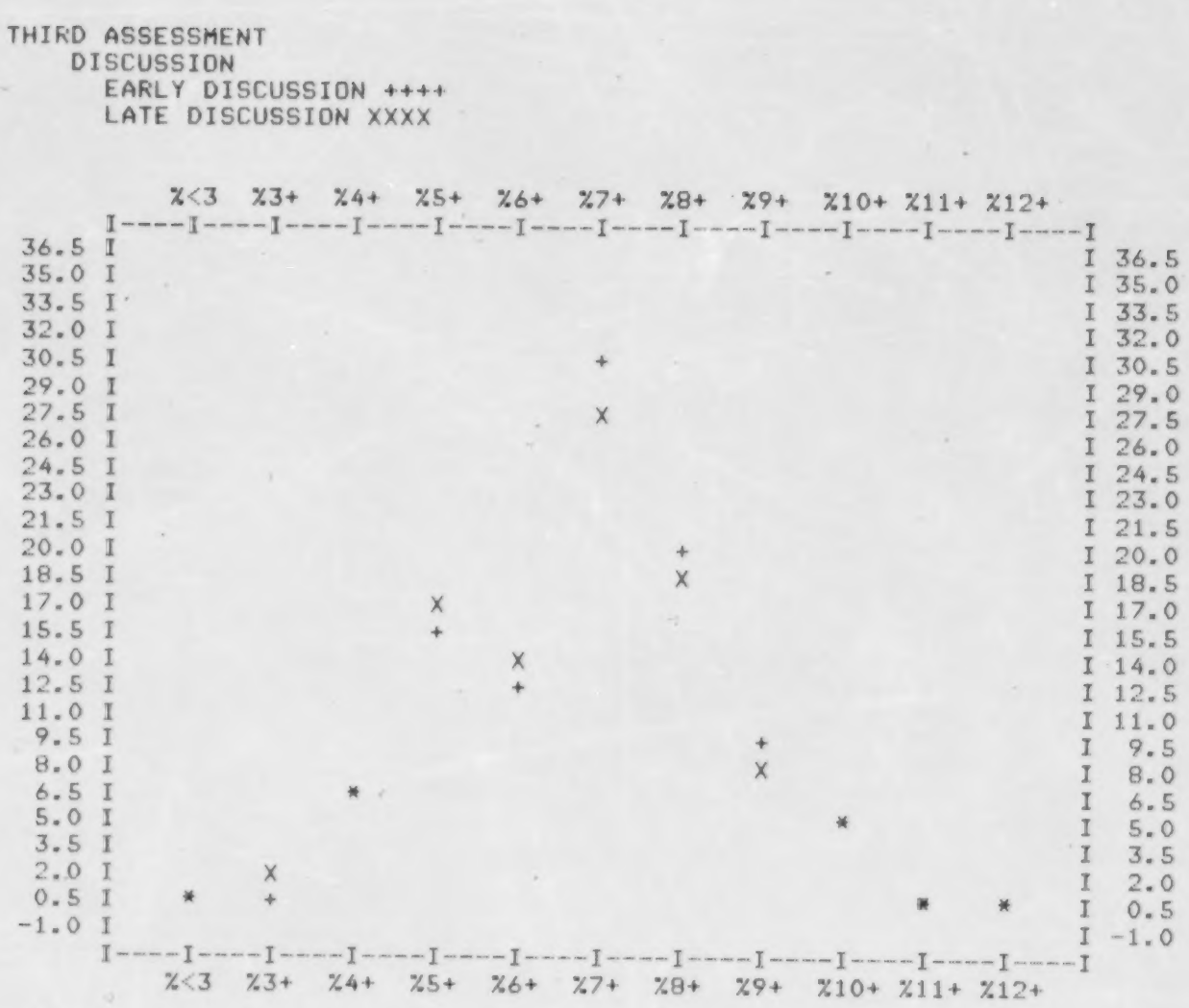
25.750	6	16.315	6	*	1	4.281	4.813
18.500	5	15.734	5	*	2	7.356	8.030
15.375	4	14.492	4	*	3	9.581	14.337
9.581	3	14.337	3	*	4	15.375	14.492
9.188	7	10.159	7	*	5	18.500	15.734
7.356	2	8.030	2	*	6	25.750	16.315
4.688	8	7.262	8	*	7	9.188	10.159
4.281	1	4.813	1	*	8	4.688	7.262
2.625	9	4.702	9	*	9	2.625	4.702
2.563	10	2.594	10	*	10	2.563	2.594
0.094	11	1.563	11	*	11	0.094	1.563

FIGURE 10



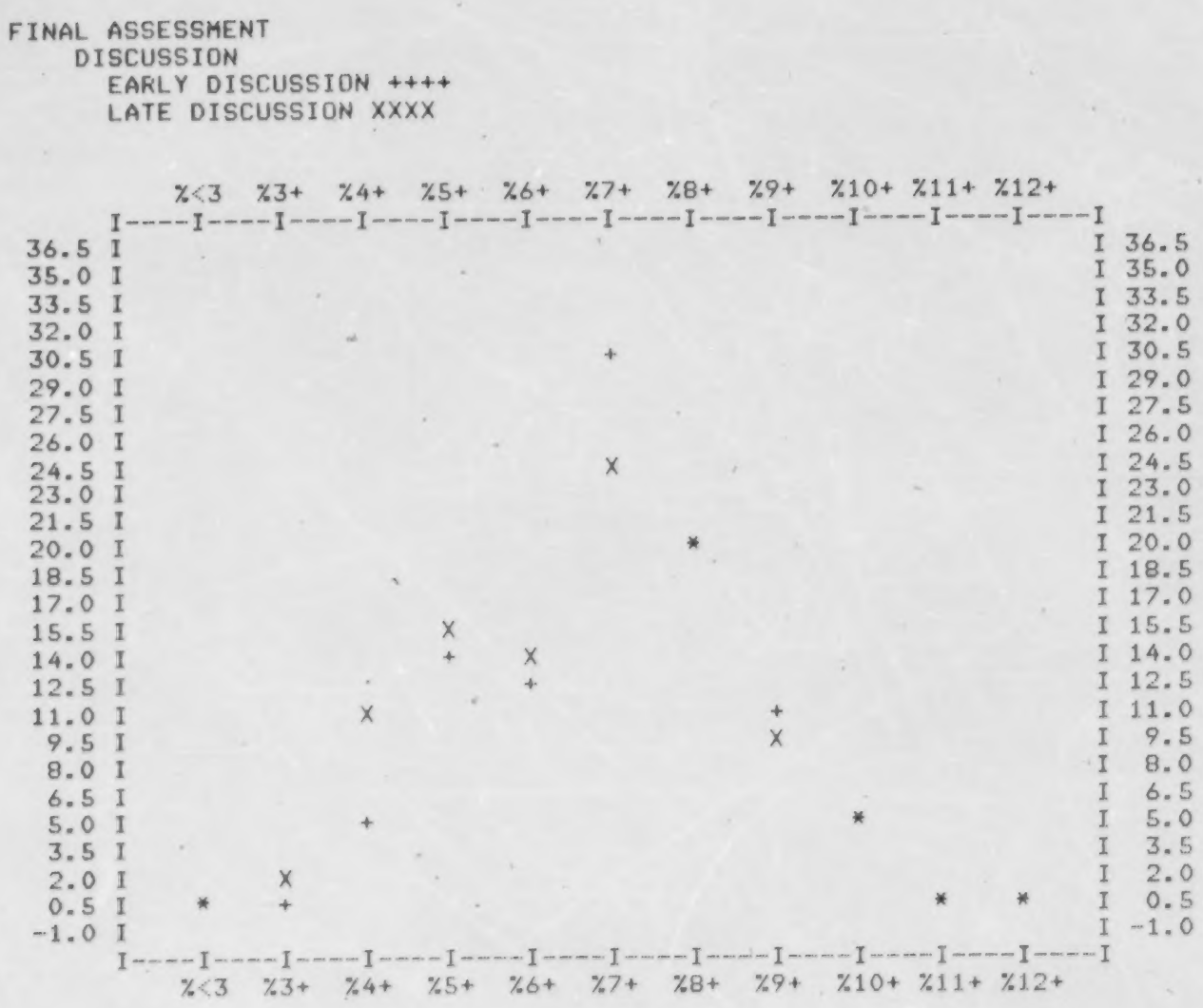
29.515	6	24.378	6	*	1	0.194	0.813
19.766	7	18.324	7	*	2	1.094	1.667
14.603	4	14.637	4	*	3	6.216	9.004
13.138	5	13.898	5	*	4	14.603	14.637
9.911	8	10.346	8	*	5	13.138	13.898
6.216	3	9.004	3	*	6	29.515	24.378
4.870	9	5.800	9	*	7	19.766	18.324
1.094	2	1.667	2	*	8	9.911	10.346
0.625	10	0.813	1	*	9	4.870	5.800
0.194	1	0.722	10	*	10	0.625	0.722
0.069	11	0.410	11	*	11	0.069	0.410

FIGURE 11



30.563	6	27.625	6	*	1	0.069	0.188
20.438	7	18.875	7	*	2	0.469	1.313
15.469	4	16.375	4	*	3	5.963	6.500
12.313	5	14.125	5	*	4	15.469	16.375
9.969	8	8.563	8	*	5	12.313	14.125
5.963	3	6.500	3	*	6	30.563	27.625
4.494	9	5.563	9	*	7	20.438	18.875
0.469	2	1.313	2	*	8	9.969	8.563
0.250	10	0.563	10	*	9	4.494	5.563
0.069	1	0.313	11	*	10	0.250	0.563
0.006	11	0.188	1	*	11	0.006	0.313

FIGURE 12



30.938	6	23.926	6	*	1	0.225	0.188
20.188	7	19.702	7	*	2	1.094	1.375
13.719	4	15.247	4	*	3	5.275	10.289
13.125	5	14.241	5	*	4	13.719	15.247
10.406	8	10.289	3	*	5	13.125	14.241
5.275	3	9.232	8	*	6	30.938	23.926
4.650	9	4.926	9	*	7	20.188	19.702
1.094	2	1.375	2	*	8	10.406	9.232
0.313	10	0.688	10	*	9	4.650	4.926
0.225	1	0.188	1	*	10	0.313	0.688
0.069	11	0.188	11	*	11	0.069	0.188

more detailed analyses confirm that the late discussion tended to lower the curve and make it more dispersed. Incidentally, the pattern in Table 3 for the early sharing teachers is also consistent with the conclusion that the sharing of information tends to lower the mean of the assessed disruption, and to increase its dispersion. However, these conclusions, like those concerning the format, involve fairly small effects and must be regarded as tentative.

In addition to considering what impact the early versus late sharing of information has on the typical location and shape of the assessed distribution, we also were interested in whether the early versus late sharing of information promotes a consensus on the location and shape of the distribution. Table 5 presents a comparison of the individual

- - - - - - - - - -

Table 5 About Here

- - - - - - - - - -

standard deviations between teachers, at each of the four time points, and separately for the teachers in the early sharing treatment versus those in the late sharing treatment. So, for example, the top panel (Panel A) of Table 5 indicates how much dissensus there was among the 16 teachers in the early discussion condition, at each time point, as to the mean of their assessment curve, and compares this with the amount of dissensus, also at each time point, among the 16 teachers in the late sharing condition. Thus, the numbers in the table reflect more consensus if they are small. At each time point, Panel A indicates that the teachers in the early sharing condition had a greater consensus as to the location of their distribution. The differences, though consistent, are not large.

Table 5

Early Information Sharing vs. Late Information Sharing
Comparison of Individual Standard Deviations

A. Index of dissensus about location

treatment		Initial	After info #1	After info #2	Final
* overall treatment	early sharing	0.906	0.380	0.205	0.244
	late sharing	1.011	0.458	0.369	0.273
within session	early sharing	0.868	0.350	0.180	0.149
	late sharing	1.007	0.418	0.373	0.227

B. Index of dissensus about apread

treatment		Initial	After info #1	After info #2	Final
overall treatment	early sharing	0.518	0.177	0.113	0.162
	late sharing	0.555	0.257	0.246	0.255
within session	early sharing	0.543	0.187	0.115	0.166
	late sharing	0.497	0.258	0.240	0.260

* NOTE: "Overall Treatment" refers to the total sigma for the 16 teachers in this treatment. "Within Session" refers to the average within session sigma (averaged over four sessions, each with four teachers). The total number of sessions includes two replications, and there are two groups of four teachers each within a replication.

Similarly, the entries in Panel B of the table indicate how much dissensus appeared as to the spread of the assessment curves, for each treatment group, and at each of the four times. Here too the teachers in the early sharing treatment have consistently greater consensus at each time. These differences, especially at times 2, 3, and 4, are fairly large. Also, the amount of consensus, on both location and spread of the assessment curves, increases greatly from time 1 to time 2 in both treatment groups. Thus, 8 out of 8 comparisons indicate that the early interaction promotes a consensus among the teachers. This conclusion is the same whether we consider the measure of dissensus to be the overall standard deviation among all 16 teachers in the same discussion treatment, or instead use as the index of dissensus the pooled standard deviation within each session within each treatment.

Tables 6A and 6B present a more detailed comparison of these relative amounts of dissensus by providing the inter-teacher standard deviations in each of the 11 score categories, at each of the 4 time points, for the teachers in the early sharing condition and for those in the late sharing condition. Table 6A gives these indexes when all teachers in a treatment are combined and treated as a single group of 16 persons. Table 6B gives the corresponding comparisons where the separate sessions are used and the index of dissensus is averaged across sessions.

- - - - - - - - - - - - - - -

Tables 6A and 6B About Here

- - - - - - - - - - - - - - -

The pattern of differences in Table 6A, though somewhat mixed, is clearly consistent with those of Table 5. At time 1, there are 7 of 11

Table 6A

Early Sharing vs. Late Sharing
Comparison of Individual Standard Deviations

		3.0 -	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-9.9	10-10.9	11-11.9	12 +
Initial Assessment												
overall	early sharing	6.006	7.473	7.476	6.702	9.223	16.303	5.765	3.439	3.914	6.207	0.272
	late sharing	7.565	6.447	10.197	7.773	9.345	11.858	5.392	5.710	5.671	4.240	3.651
After info #1												
overall	early sharing	0.401	1.828	5.014	4.215	5.090	8.022	9.366	3.529	3.818	1.133	0.250
	late sharing	1.515	2.280	5.641	5.428	7.095	7.681	9.853	4.296	3.902	1.517	1.395
After info #2												
overall	early sharing	0.250	1.087	3.326	4.129	6.258	5.112	6.022	4.084	2.555	0.548	0.025
	late sharing	0.750	2.469	3.033	3.557	6.642	4.801	9.237	3.204	3.705	1.413	1.250
Final Assessment												
overall	early sharing	0.545	1.676	2.537	4.697	7.032	7.620	5.659	3.313	2.896	0.574	0.250
	late sharing	0.750	1.996	5.606	4.103	5.307	8.933	7.171	2.999	2.844	1.401	0.750

Table 6B

Early Sharing vs. Late Sharing
Comparison of Individual Standard Deviations

		3.0 -	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-9.9	10-10.9	11-11.9	12 +
	Initial Assessment											
within	early sharing	5.950	8.155	7.692	5.563	9.679	17.168	4.864	3.169	3.180	4.395	0.188
	late sharing	5.820	5.834	9.861	6.345	9.924	9.632	4.952	6.104	4.684	3.061	2.470
	After info #1											
within	early sharing	0.388	1.342	4.274	4.376	4.607	7.681	9.550	3.627	3.629	0.903	0.138
	late sharing	1.479	2.321	5.089	5.627	6.958	7.547	9.226	3.264	3.571	1.142	0.820
	After info #2											
within	early sharing	0.138	0.666	2.894	4.086	6.129	5.417	5.894	3.903	1.950	0.424	0.013
	late sharing	0.375	2.470	2.678	3.344	5.233	4.243	9.530	3.010	3.545	1.125	0.625
	Final Assessment											
within	early sharing	0.434	1.355	1.623	3.866	5.586	7.076	5.212	2.861	2.125	0.549	0.138
	late sharing	0.375	1.876	5.961	4.322	5.095	8.090	5.839	1.922	3.022	0.910	0.375

comparisons in the direction of greater consensus among the early sharing teachers. The corresponding figures at times 2, 3, and 4 are: 10 of 11, 7 of 11, and 7 of 11. Thus, all 31 of 44 comparisons are in the direction of greater consensus among the early sharing teachers. Although the 44 comparisons are not strictly independent, it is interesting to note that the chance of such a pattern (31 of 44) occurring if the true probability were .5, and if the trials were independent, is .005.

The results in Table 6B are again generally consistent with the idea that there is greater consensus among the teachers in the early sharing treatment. Of 11 comparisons possible at each time point, the majority always indicate that the early sharing groups have greater consensus. At time 1, it is 7 of 11; at time 2, also 7 of 11; at time 3, 6 of 11; and at time 4, 8 of 11. This makes a total of 28 of 44 comparisons which are in the direction indicating greater consensus for the groups using early sharing.

One incidental factor which we decided to examine was the ordering of instruction in the two assessment methods. That is, for sessions 1 through 4, in giving instructions about making the assessment curves, we explained the density method first, and then the cumulative method. For sessions 5 through 8, the ordering of the instruction was reversed, so that the cumulative assessment method was explained first. These comparisons were done as a check on one possible source of distortion. Notice that this is also a comparison of the early sessions against the later sessions. The results of these comparisons are clear. Not one of 8 comparisons of distribution means is significant, and 7 of the 8 are

quite small. This indicates (as expected) that there is no effect of the different ordering of instruction.

5.3 Impacts of Evidence Patterns

In addition to studying possible effects of the method of assessment and the timing of the group information sharing, we were interested also in the response by the teachers when they received particular patterns of evidence. As described earlier, the evidence inputs differed in level and in the perceived source. One set of evidence indicated relatively low achievement; the other set indicated relatively high achievement. Also, some teachers received the low evidence with a description stating that it resulted from a sample testing, and others received the same evidence with a description stating that it was the judgments of teacher-experts like themselves. Finally, the sequence of delivery of the evidence also was varied across teachers. Some received the low evidence first and others received the high evidence first. Some received the sample evidence first, and others received the expert evidence first.

Most central for this report is the question of the relative weighting given to the two different sources. However, we shall also discuss briefly the impacts of other configurations of evidence as well.

To carry out this analysis, we can use the distributions assessed at each time, as was done for the earlier comparisons. In addition, comparisons also are made on the changes between the various assessments in response to each new evidence item. The assessment curve means and standard deviations for teachers exposed to each evidence pattern at each time point are shown in Table 7. The interpretation of this table

- - - - - - - - - -

Table 7 About Here

- - - - - - - - - -

is somewhat complicated. First, the assessments at time 1 logically cannot show any impact, but rather show the closeness among the four groups. In terms of the F-statistics and the inspection of cell entires, it appears that the randomization of groups was quite successful with respect to assessment curve location (mean), but that the SL&EH subset of teachers differed from the other three subsets with respect to the assessment curve spread (standard deviation). Thus, comparisons of the assessment curve levels at later times may be made without fear of distortion, but comparisons of assessment curve dispersions at later times should take into account the initial difference between the SL&EH subset and the other subsets.

Inspection of the column labeled "After info #1" shows some interesting patterns. First, all four groups move their location upward, on average a distance of 0.749 grade-equivalent units. This indicates that the teachers initially judged the achievement levels to be somewhat lower than they were in fact at the Daw School. It also shows that the teachers as a group adjusted their assessment curve levels substantially and quickly in response to the quantitative information given in the first wave of evidence. A second feature of the assessment curve locations at time 2 is that the two lower means (7.134 and 7.275) occur in response to the "low" evidence, and in each case represent a shift of approximately 0.7 units from the original position. Thus, it appears that this parti-

Table 7

Evidence Pattern Comparisons
Comparison of Means and Sigmas

means

treatment	Initial	After info #1	After info #2	Final
SL&EH	6.437	7.134	7.356	7.363
EH&SL	6.687	7.309	7.335	7.244
SH&EL	6.698	7.705	7.497	7.414
EL&SH	6.606	7.275	7.321	7.364
F-statistic	0.119	3.400	0.573	0.529
Average	6.607	7.356	7.377	7.346

sigmas

treatment	Initial	After info #1	After info #2	Final
SL&EH	1.500	1.507	1.460	1.554
EH&SL	2.026	1.723	1.651	1.677
SH&EL	1.928	1.766	1.676	1.682
EL&SH	2.116	1.675	1.561	1.639
F-statistic	2.198	2.081	2.330	0.535
Average	1.892	1.668	1.587	1.638

cular evidence was regarded as equally impressive whether it was labeled as coming from the sampling of students or from expert judgments. The response in the two subsets which received "high" evidence in the first wave of evidence is less simple. Among the teachers who received high evidence labeled as coming from an expert judgment, the distance changed, on average, was 0.622, which is comparable to the other changes, but perhaps less than might be expected in view of the greater discrepancy between the content of the evidence and the initial judgments. Among the teachers who received the same "high" evidence, but had it labeled as the result of a sample testing, the distance moved was 1.007. This is more congruent with the relatively greater discrepancy between the teachers' average initial judgment and the achievement level indicated by the evidence. These results suggest that when an item of evidence is somewhat surprising, greater weight will be attached to evidence from an objective method than to evidence obtained via expert jdugment. Notice that in the two cases where the evidence was less discrepant with the initial judgments (i.e., less surprising) the two sources were equally important.

Examination of the standard deviations at time 2 likewise indicates a general decline in assessment curve spread. This too is consistent with the content of the evidence. However, the teachers who received SL evidence at the first wave did not decrease their assessed spread. This is further evidence that the teachers are responding to the information in an intelligent manner. Among the teachers who had overestimated the spread, there was a decline in assessed spread, but in the subset of teachers whose estimate had been slightly too small (the true

spread index is about 1.63), no decline occurred.

The third column (time 3) shows a considerable convergence of the assessment curve means. This is as would be expected, because all teachers have received the same evidence content by this time. In other words, although the labeling of the two sets of evidence varied, and the sequence of delivery varied, each teacher eventually did receive the same total evidence content. The two groups which received the "high" evidence as the second evidence each raised their means slightly (to 7.356 and to 7.321). The groups which received "low" evidence as the second evidence continued to move toward a figure of about 7.42, which is very close to the true location. The F-statistic indicates that the four subgroups here are indistinguishable. This is as would be expected if teachers are efficient processors of information.

The pattern among the standard deviations at time 3 is less clear, but shows a continued shrinkage among the three subsets which initially overestimated the amount of spread.

The only input to the teachers between time 3 and time 4 was feedback about the ideas of the other teachers in the working group. For those sessions which had information sharing early, the feedback between time 3 and time 4 was a report of the average distribution assessed at time 3. For the groups which had information sharing later, this information sharing (a presentation of the factors each teacher individually felt to be relevant to making an assessment), as well as the feedback of the average distribution, were events occurring between assessment 3 and assessment 4. As can be seen from an inspection of column 4 of Table 7,

there is little movement of the assessment curve location between these two sessions. The assessed spreads at time 4 are somewhat more alike across subsets than were the earlier spreads, and this is caused mainly by a movement upward of the two spreads which were somewhat small. This too indicates the responsiveness of the teachers as a group in adapting their judgments.

To summarize the interpretation of Table 7, we conclude that the teachers treat the information inputted to them as important, and they attach nearly as much weight to information from the judgments of other experts as they do to information obtained from an objective testing. Their judgments move quite closely in response to the content of the information they receive. When the received information is surprising, there is evidence of a tendency to give more importance to information from an objective testing than to information from other expert judgments. This is a refinement on the general conservative tendency noted by Edwards.

Table 8 shows the differences among the four subsets when the assessment curves are represented by the 11 percentages assigned to each grade-equivalent category. Not surprisingly, the major differences among the four subsets appear at time 2, after receipt of evidence #1.

- - - - - - - - - - -

Table 8 About Here

- - - - - - - - - - -

To see the general import of Table 8, it is perhaps sufficient to examine the second panel (labeled "after info #1"). In particular, the column for scores of 5.0-5.9 shows a "significant" difference among

Table 8

Comparison of Percentage Assignments under Different Evidence Patterns

Initial Assessment	3.0 -	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-9.9	10-10.9	11-11.9	12 +
SL&EH	4.813	5.250	9.375	18.250	18.875	29.125	8.438	4.562	0.750	0.375	0.188
EH&SL	4.000	8.147	13.711	18.109	14.592	11.880	12.755	6.587	4.717	4.375	1.125
SH&EL	4.625	6.875	14.000	12.000	14.375	25.500	9.000	5.250	4.063	2.438	1.875
EL&SH	4.750	10.500	10.750	11.375	20.625	17.625	8.500	7.500	5.125	3.125	0.125
	0.022	0.816	0.468	2.535	0.906	2.518	1.144	0.575	1.363	0.804	0.774
After info #1											
SL&EH	0.013	0.438	8.550	17.938	15.250	30.500	15.750	8.813	2.425	0.313	0.013
EH&SL	1.125	1.876	9.001	12.986	14.364	22.606	21.583	9.981	6.103	0.375	0.000
SH&EL	0.750	1.500	5.569	10.958	11.444	23.083	25.653	11.333	7.632	1.257	0.819
EL&SH	0.125	1.708	7.319	16.597	13.014	31.597	13.194	10.389	5.181	0.750	0.125
	1.866	0.778	0.614	5.026	0.571	3.347	3.506	0.557	3.162	0.856	1.250
After info #2											
SL&EH	0.013	0.438	6.175	14.938	14.000	30.000	23.250	7.188	3.800	0.188	0.013
EH&SL	0.375	0.625	6.750	17.000	14.750	27.750	16.000	9.750	6.625	0.375	0.000
SH&EL	0.125	1.125	5.875	15.625	10.750	28.500	20.000	11.500	5.188	0.688	0.625
EL&SH	0.000	1.375	6.125	16.125	13.375	30.125	19.375	8.625	4.500	0.375	0.000
	0.774	0.383	0.101	0.389	0.563	0.390	1.212	2.184	1.171	0.282	0.987
Final Assessment											
SH&EL	0.263	1.438	6.175	12.563	16.000	30.000	19.000	10.188	3.925	0.438	0.013
EL&SH	0.375	1.000	9.577	16.244	12.732	27.351	18.030	8.214	6.101	0.375	0.000
SH&EL	0.188	1.250	6.625	15.250	13.625	26.250	20.375	10.250	4.750	0.938	0.500
EL&SH	0.000	1.250	8.750	13.875	12.375	26.125	22.375	10.625	4.375	0.250	0.000
	0.452	0.071	0.857	1.071	0.537	0.304	0.681	0.938	0.870	0.613	1.720

the four groups (F = 5.026). This difference occurs because the two subsets which received low evidence allocated a greater percentage to this category (i.e., 17.938 and 16.597) than did the subsets of teachers who received high evidence at wave #1 (i.e., 12.986 and 10.958). Similarly, at the same time, the columns of percentages for scores from 7.0-7.9 and from 8.0-8.9 each show "significant" F-statistics, and each is accounted for by the level of the evidence received at that time.

Because there is some risk that the average changes described in Tables 9 and 10 may conceal the individual dynamics of revision of assessment, we carried out a further analysis in which each individual teacher's revision of assessment was calculated in light of the particular evidence just received by that teacher and the teacher's own immediately preceding assessment. In computing this index of revision, we scored a shift as positive if it reflected a movement in the direction implied by the evidence and as negative if it reflected a movement in the direction opposite to that implied by the evidence.

The revision index scores were computed separately for each teacher and then aggregated across all teachers at a given time, keeping separate those teachers who had just received "sample" evidence from those who had just received "expert" evidence. The average revisions observed between time 1 and time 2, time 2 and time 3, and a combined amount of revision collapsing both of the changes are presented in Table 9. Table

- - - - - - - - - - -

Table 9 About Here

- - - - - - - - - - -

9 shows that the revision after sample input is consistently a larger

Table 9

Source of Evidence	Change T2-T1	Change T3-T2	Combined Change
sample	1.004 (16)	0.204 (16)	0.562 (32)
expert	0.758 (16)	0.095 (16)	0.420 (32)
F-ratio	0.870	0.768	0.658
% E/S	75.53	46.79	74.72

change than the revision after the expert source. For the combined changes, the impact of the expert source is 74.72% as large as the impact of the sample source. These results are indicative of a tendency to attach greater weight to the sample evidence than to the expert evidence. However, they still indicate the strong impact of the expert evidence on the judgments made by these teachers. In other words, the teachers in this study tended to attach substantial credence to the considered judgments of other teachers like themselves, when such judgments were made about a concrete educational situation.

As an additional step in analyzing these indexes of revision, we calculated also the revisions made by each teacher in each of the percentage assignments, and compared these for the teachers receiving evidence from a sample with the same assignments for teachers receiving evidence from an expert judgment. The results of this comparison, for each of the revisions separately and for the combined revision including both time intervals, are given in Table 10. A detailed recapitulation

- - - - - - - - - - -

Table 10 About Here

- - - - - - - - - - -

of the contents of this Table would be too lengthy. In general, the results confirm those of Table 9, and indicate that there is a somewhat greater revision of assessment in response to the sample evidence than in response to the expert evidence. Again, however, perhaps the most interesting outcome is that the teachers do revise their opinions substantially in response to the expert judgments of other teachers.

Table 10

Change T2 - T1	3.0 -	3-3.9	4-4.9	5-5.9	6-6.9	7-7.9	8-8.9	9-9.9	10-10.9	11-11.9	12 +
sample (16)	4.338	5.094	5.053	3.823	8.722	8.479	11.858	7.042	4.428	0.622	0.615
expert (16)	3.750	7.531	4.533	5.173	5.633	14.599	6.761	4.282	1.790	3.188	0.563
F-ratio	0.064	1.087	0.024	0.495	0.961	4.025	1.690	2.417	4.659	2.267	0.007
Change T3 - T2											
sample (16)	0.438	0.792	2.222	4.021	5.237	3.308	4.507	2.237	3.077	0.188	0.063
expert (16)	0.313	0.188	2.410	3.833	1.778	3.708	6.576	0.229	1.910	0.347	0.097
F-ratio	0.112	1.251	0.016	0.019	3.909	0.019	0.576	2.043	0.766	0.309	0.138
Combined Change											
sample (32)	2.388	2.943	3.637	3.922	6.980	5.894	8.182	4.639	3.752	0.405	0.339
expert (32)	2.031	3.859	3.471	4.503	3.706	9.154	6.669	2.256	1.850	1.767	0.330
F-ratio	0.081	0.425	0.008	0.248	3.189	1.958	0.385	3.646	4.490	2.353	0.001

5.4 Overall Accuracy of the Teacher Assessments

The general question of the accuracy of the teacher's assessments can be addressed by comparing the "actual" distribution of mathematics skills among the 309 7th graders at the Daw School with the assessed distributions. Table 11 presents the results of the comparisons at each of the four time points. The table shows the actual percentages in

- - - - - - - - - - -

Table 11 About Here

- - - - - - - - - - -

each of the grade-equivalent intervals, and the percentages judged to be in these same intervals by the 32 teachers in the study. The Table also shows the actual mean and actual sigma of the Daw School mathematics skills, plus the mean and sigma of the assessed distribution.

At time 1, there is not a very close correspondence between the actual distribution and the assessments made by the teachers. The actual mean of the 309 students is 7.409, but the mean of the assessed distributions is somewhat lower, at 6.607. This indicates that the teachers in this study tended to generalize their own personal experience in a large-city system to the Daw School context. Comments by various teachers during the sessions confirm this interpretation. Also, the assessed distribution at time 1 is somewhat more widely spread than the true distribution (actual sigma = 1.632, assessed sigma at time 1 = 1.892). A Kolmogorov-Smirnov test of the similarity of the two distributions yields a D-statistic of 21.622, with a p-value less than .10 if the two distributions are the same.

However, at time 2, after the first wave of quantitative evidence

Table 11

Overall Accuracy of Assessments

Grade-Equivalent	Daw	Time 1	Time 2	Time 3	Time 4
3.0 -	0.000	4.547	0.503	0.128	0.206
3-3.9	1.942	7.693	1.381	0.891	1.234
4-4.9	5.825	11.959	7.610	6.231	7.782
5-5.9	12.298	14.933	14.620	15.922	14.483
6-6.9	14.563	17.117	13.518	13.219	13.683
7-7.9	24.919	21.033	26.947	29.094	27.432
8-8.9	22.654	9.673	19.045	19.656	19.945
9-9.9	11.651	5.975	10.129	9.266	9.619
10-10.9	5.825	3.664	5.335	5.028	4.788
11-11.9	0.324	2.578	0.674	0.406	0.500
12 +	0.000	0.828	0.239	0.159	0.128
mean	7.409	6.607	7.356	7.377	7.346
sigma	1.632	1.892	1.668	1.587	1.638
(K-S) D =		21.622	5.031	5.938	5.273
probab greater than		.10	.20	.20	.20

has been received, the assessed distribution becomes remarkably similar to the actual distribution. Whereas the actual mean achievement is 7.409. the assessed mean becomes 7.356; similarly, the actual sigma of the distribution is 1.632, and the assessed sigma is now 1.668. The entries in the individual intervals also resemble quite closely the actual distribution, with one notable difference. The actual distribution of all 309 scores is unimodal, but the assessed distribution at time 2 (and at time 3 and time 4) is bimodal, with a second small peak in the interval from 5.0 through 5.9. This discrepancy between the actual distribution and the assessed distribution is not a sign of imprecision among the teachers, however. On the contrary, the evidence they received was constructed so that just such a second peak was presented in each of the evidence inputs. Thus, the presence of this bimodality in the assessed distributions indiates that, on the average, the teachers' assessments follow very faithfully the quantitative evidence they have received. Also, despite the bimodality, the distributions assessed after time 1 all are quite close to the actual distribution. The Kolmogorov-Smirnow D-statistice are all between 5.0 and 6.0 , and indicate no statistically significant differences between the distributions assessed and the actual distribution.

6. Conclusions

Several kinds of conslusions may be drawn from the results of this study. First, the general procedures used in the study are viable ones. The teachers participated effectively and with a negligible amount of

difficulty in the tasks they were asked to do. Impressionistically, we judge that the session was somewhat lengthy, and that there was much listening for the teachers. However, these characteristics can be avoided in subsequent projects simply by scheduling a series of several related sessions and specifically devoting some of these to orientation and training.

In relation to the particular experimental factors studied in this investigation, several results have emerged. First, there appears to be no difference or only slight differences between the two methods of assessment representation (i.e., the density method versus the cumulative method). At the first wave of assessments, the cumulative method tends to rise somewhat too quickly. This appears to be an artifact of the tendency to begin at the left edge and proceed sequentially from left to right. In this respect, the cumulative method is less desirable. However, in the later assessments, the cumulative method shows no differences from the density method. This indicates that either method may be used in research studies, and that the results from either method are (at least for most purposes) directly comparable. A further implication, consistent with the preceding section on results, is that the teachers who served as the judges learned quite quickly to adjust their curves to reflect the evidence of the curves presented to them.

With respect to the effects of the timing of the information sharing activity, it appears there is a tendency for the early sharing to cause a narrowing of the range of assessments, and a piling-up of scores in the 7.0-7.9 interval, in comparison to the assessments made by the teachers

in groups experiencing the later sharing. This narrowing may or may not be viewed as desirable. Empirically, it is somewhat more accurate in this study, although the actual scores are somewhat higher.

With regard to the weighting of sample evidence versus expert evidence, the results indicate that both kinds of evidence are weighted sizeably, with the sample evidence being given more weight. It also appears likely that this differential importance tendency is more pronounced when the evidence is surprising--that is, when it is not similar to the teacher's previous beliefs. One interesting aspect of these results is the sizeable revisions made in response to either kind of evidence. This group of teachers takes research evidence quite seriously when that evidence concerns a concrete situation with which they are involved. These teachers also are willing to place confidence in the expert judgments of others like themselves. This confidence is less than, but comparable to, the confidence they attach to evidence obtained via an objective testing.

Overall, the study provides clear evidence that the teachers <u>on the average</u> perceive the information quite accurately. If anything, they accept such quantitative data somewhat too readily--even incorporating the artificial bimodality of the evidence into their assessments. This suggests a conclusion quite different from that implied by Rosenthal and Jacobson (1968), who regard teachers as being prone to stereotyping. In our study, the teachers were sensitive both to objective evidence and to subjective evidence when both were presented in a clear and quantitative form. They recognized the existence of extreme scores, and in fact

tended to expect somewhat more spread than our Daw School data contain. The rapidity of their reaction to the evidence, indicated by the fact that the assessments at time 2 already correspond quite closely to the actual distribution, by the large revisions occurring at time 2, and by the fact that those revisions are entirely congruent with the content of the evidence each teacher has received, shows a tendency to respond in a quick and objective manner to new data on the performance of a student or a group of students. Of course, a critic might maintain that the behavior we observed shows simply that the teachers are ready to place a great deal of faith in test scores. We would point out, however, that the teachers here placed a nearly equivalent degree of faith in evidence which was described to them as being based on the judgments of other teachers.

This pattern of reasonableness and good assessments in the aggregate, however, masks substantial divergences in the assessments made by individual teachers. That is, the individual curves drawn by a few teachers depart quite substantially from the aggregate curve, or from the evidence they have received. Impressionistic analysis of the verbal comments made by the teachers during the sessions indicates that there is also a wide degree of variability in the preferred strategies of assessment, in the kinds of information used by different teachers, and in the weightings they attach to various evidence sources.

Thus, our conclusion on this point is a double-edged one. The assessments made by teachers are, in the aggregate, of remarkably high quality; these same assessments, considered individually, are marked by considerable

idiosyncrasy.

As mentioned previously, our results show relatively slight effects of any incidental features of the task. The method used to report the assessment, the sequence of evidence, and the linking of level and source, all have no major impact on assessments. This is a highly promising result for the practical future of techniques such as these. It indicates that they are feasible and flexible.

A major practical conclusion of the study, though it should be further explored, is that it is feasible, justifiable, and desirable to use groups of teachers as a means of developing diagnoses and assessments about concrete aspects of the school situation. The pooling of judgments allows us to capitalize on teachers' collective professional wisdom and to neutralize their individual idiosyncrasies. This conclusion is consistent with the arguments proposed by the advocates of the so-called "delphi" technique. Our work makes clear, however, that the advantage comes less from the particular group technique used than from the systematic pooling of the members' judgments. Another view on this would be to regard the assessments made by each teacher as analogous to a sample. If the sample is tiny, it is likely to be somewhat far from the mark. However, even a modest size sample can provide great benefits in comparison to a single case. This same logic seems to apply to the judgments proffered by teachers.

Naturally, there are many further questions to be addressed. In the data already collected, further analysis is needed to trace out in more detail the processes by which teachers formulate their assessments. Another line of analysis would explore individual characteristics that

may be associated with different patterns of information processing. At a very concrete level, further studies are needed to extend the results beyond the junior high school mathematics context. At the theoretical level, work is needed to study the determinants of "acceptability" as well as of accuracy. Yet, despite the size of this agenda, the results presented here provide strong grounds for developing this general body of techniques. They allow school management to build upon a highly valuable but neglected resource--the considered professional judgments of the teachers in a school.

FOOTNOTES

1. One exception, published since the data in the CAP study was collected, is a report by Richard Shavelson and others on teachers' sensitivity to the reliability of information (Shavelson et al., 1977).

References

Collet, Lee. FEHR-Practicum. Final Report to the U.S. Office of Education, Ann Arbor, Michigan, the University of Michigan, 1974.

Edwards, Ward. Conservatism in Human Information Processing. In Benjamin Kleinmuntz, ed., Formal Representation of Human Judgment. New York: John Wiley, 1968.

Fennessey, James. Focused Flexibility in a Secondary School: A Description and Analysis. Baltimore, Maryland: The Johns Hopkins University, Center for Social Organization of Schools, Report No. 149. 197

Fennessey, James. Using Achievement Growth to Analyze Educational Programs. Baltimore, Maryland: The Johns Hopkins University, Center for Social Organization of Schools, Report No. 151, 1973.

Fennessey, James. Improving Inference for Social Research and Social Policy: The Bayesian Paradigm. Social Science Research, 1977, 6, 309-327.

Karweit, N.L. Rainy Days and Mondays: An Analysis of Factors Related to Absence from School. Baltimore, Maryland: The Johns Hopkins University, Center for Social Organization of Schools, Report No. 162, 1973.

Richards, James M. A Simulation Study Comparing Procedures for Assessing Individual Educational Growth. Journal of Educational Psychology, 1976, 68, 603-612.

Rosenthal, Robert and Lenore Jacobson. Pygmalion in the Classroom. New York: Holt, Rinehart and Winston, 1968.

Shavelson, Richard J., Joel Caldwell and Tonia Izu. Teachers' Sensitivity to the Reliability of Information in Making Pedagogical Decisions. American Educational Research Journal, 1977, 14, 83-97.

Shulman, Lee and Arthur S. Elstein. Studies of Problem Solving, Judgment, and Decision Making: Implications for Educational Research. In Fred N. Kerlinger, ed., Review of Educational Research, Vol. III. Itasca, Illinois: F. E. Peacock, 1975.

Tversky, Amos, and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. Science, September 1974, 185, 1124-1131.

Appendix A

Copy of Daw School Descriptive Booklet

Section I: GRANDVIEW AND ITS EDUCATIONAL SYSTEM

PROFILE OF GRANDVIEW

INTRODUCTION: Grandview is a city of 31 square miles in which 122,594 residents live.* Geographically, the city is characterized by two rivers which flow through it. The North River runs through the heart of the industrial district and the South River through the municipal area. Both rivers flow toward Grandview from the surrounding countryside and finally converge east of the city. (See the City Map on the next page).

THE PEOPLE AND QUARTERS: The residents of Grandview represent a wide variety of ethnic groups. The population is 62% white and 38% non-white; the non-white proportion of the population has been increasing slowly over the past 20 years. Grandview, like all large cities, has citizens belonging to many socio-economic levels; this variety is reflected in its housing. The homes in the city range from modern high-rise apartment complexes near the new shopping mall, through lovely old homes found in the center of the city, through the more modest middle income dwellings in most of the city, to the large but dilapidated homes on the south side of the city.

WORK AND CONSUMPTION: More than 50,000 residents of Grandview have jobs in the city in a wide variety of occupations ranging from clerical and service jobs to research work associated with State University. Some 17,000 of these workers are employed in 112 manufacturing industries; average manufacturing pay rate is $3.03 per hour ($119 per week). More than 93 percent of the total work force

---

*Most of the figures are based on 1970 data, unless otherwide specified.

City Map

BY CENSUS TRACTS

is employed, giving the city an average household income of $6,943, which is higher than the national average of $4,630.

There are 747 retail and wholesale outlets, including those in 5 principal shopping areas, producing $150,185,000 in retail sales.

TRANSPORATION AND COMMUNICATION: A variety of public transportation facilities is provided for the residents of Grandview. People can visit other cities through one of two railroads crossing the city, or in an airplane leaving Burns County Airport. Around the city, four intra-city bus lines and one county line are available. Moreover, there are 35 truck lines to other principal cities throughout the country.

People enjoy various programs provided by four radio stations and two TV stations. Also, people obtain a variety of information through two newspapers, one with daily circulation of 50,000 and the other with weekly distribution of 1400.

HEALTH, WELFARE AND RECREATION: The Grandview Health Department, governed by the Board of Health, provides a variety of services to the city, such as immunization programs for children, sanitation regulation, health education, and operation of several clinics in the city. The County Welfare Department aids a variety of citizens who are in need of assistance, such as the blind, disabled, dependent children, the aged, and the unemployed.

Among medical institutions in Grandview, the Mercy Hospital has been taking a leading part in health care of the people. It was built in 1910 as one of the first teaching hospitals in the state. From that time on, it has continued to be a leader in innovative techniques and in attracting qualified staff. Today it has a bed capacity

of 545. State University has a nursing and medical internship program, which has been long connected with Mercy Hospital.

The residents of Grandview can enjoy a variety of the cultural and recreational benefits associated with metropolitan living. Their Symphony provides musical entertainment of high quality and the city's many parks and playgrounds offer space for relaxing for young and old alike. In addition, golf can be enjoyed on one of three 18-hole courses.

THE EDUCATIONAL SYSTEM IN GRANDVIEW

The Grandview school system is typical of a system which has grown steadily in recent years. It has many of the characteristics of a small system in which the community and parents are concerned and interested in the operation of the schools. Yet with the gradual increase in the population, especially in the black population, the school system has had to reorient itself.

SCHOOL POPULATION: in 1975, the school system population contained 23,883 pupils of which 36% were black (Appendix-Table 3). The working class areas had a predominantly black student population while the upper middle class areas contained few, if any, black pupils. Recently, this separation of the races has been reduced in the central sections of the city, but some work is needed to alleviate the possibility of problems in other parts of town. Several plans have been considered by Grandview to help solve this racial imbalance, but one has not been proposed which satisfies the various factions within the city.

TEACHERS: The teachers in Grandview are generally well qualified and rank above the national average in their educational attainment level and salary. This has been due largely to the strong teacher's union, which has attracted able and dedicated teachers. Their latest struggle has been to decrease the pupil-teacher ratio from an average of 28 to 1 to 23 to 1, but as yet the cost is too great for such a change.

SCHOOLS: The school system has grown by almost 72% since 1950 and now contains 21 grade schools, 7 junior high schools and 3 senior high schools (Appendix-Table 3). The school year is 39 weeks long, beginning in early September and ending in the middle of June.

The junior and senior high schools are departmentalized and children are grouped by academic achievement and ability in most areas of study. In the junior and senior high program, there are 3 curriculum groups of students: (1) college preparatory enriched, (2) college preparatory regular, and (3) occupational preparatory. Students are chosen for the enriched classes on the basis of (a) grades, (b) initiative, and (c) teacher recommendations. Pupils decide whether they wish to be in college preparatory regular or occupational preparatory programs, and counseling is available at every level to guide them in making their choice. Approximately 15% of all Grandview pupils are classified as college preparatory enriched; another 40% are regular college preparatory. The occupational preparatory contains the remaining 45% of the students. This latter program serves its students generally quite well in preparing them for occupations immediately upon high school graduation. There has, however, been some pressure to make the preparation for jobs

more relevant to Grandview. At the present time, negotiations are taking place with various companies in Grandview to place high school students in internship programs for a small portion of the senior high school students' occupational program. This would mean a great change in the educational framework, since students would physically leave the school, for a sizeable portion of time and is therefore being given careful consideration.

SPECIAL PROGRAMS: In the last ten years, Grandview has instituted special programs in some of the elementary, junior high and senior high schools for children with various physical or mental disabilities. There are programs for the physically handicapped, the mentally retarded and the emotionally disturbed.

PARENT AND COMMUNITY INVOLVEMENT: Grandview has a long tradition of community involvement and support for the school system. In past years, the electorate has supported both school budget and bond issues overwhelmingly. Recently, however, there has been some decline in support for these issues, and last year, a proposed budget increase was voted down. The city schools personnel now feel that they must form a closer relationship with the parents, especially since school costs and complexities have grown rapidly in the recent years.

A recent poll taken by State University showed that 80 percent of parents of children in Grandview schools felt their child's teacher was doing an average, or better-than-average job; 91 percent felt the teacher was qualified. Findings concerning the principals and other administrators were similar.

Certain groups of parents, especially those in black or low income areas, showed a lower degree of satisfaction. For instance, overall 39 percent of Grandview parents felt the school was not doing an adequate job in preparing their children for a job or for college; while in the black and low income areas, 52% of the people felt this same dissatisfaction with the job preparation the children were receiving in the schools.

UNIVERSITY AND COLLEGE: In Grandview, higher education is provided by State University and Grandview Community College.

The Grandview campus of State University was built in 1950 as the largest campus in the state. It is situated among the rolling hills on the edge of the city. It is a coeducational institution with approximately 20,000 full-time students. In recent years, more than a dozen new buildings have been added to the 32 acre campus. Grandview Community College was built in 1968 to meet the needs of the citizens of the town who wanted further education. Already, there are 2,500 part-time students and 250 full-time students at the college. The college offers associate degrees in chemical, mechanical and industrial technology; business, finance and remedial courses. It also offers two year courses in both the liberal arts and the sciences which may be used as transfer programs should the student decide to continue his education beyond the associate level.

APPENDIX

TABLE 1

1970 POPULATION BY AGE

AGE	TOTAL	NON-WHITE	PERCENT	WHITE	PERCENT
Under 5 years	14,009	3,781	27.0%	10,228	73.0%
5-11 yrs.	13,743	4,589	33.4%	9,154	66.6%
12-14 yrs.	6,673	2,381	35.7%	4,292	64.3%
15-17 yrs.	6,579	2,233	33.9%	4,346	66.1%
18 yrs. & over	81,590	35,574	41.1%	48,016	58.9%
Total	122,594	46,558	38.0%	76,036	62. %

TABLE 2

ECONOMIC CHARACTERISTICS OF THE POPULATION: 1970

Total Personal Income	$287,775,576
Per Capita Personal Income	2,348
Per Capita Effective Buying Income	2,008
Household Income (Average)	6,943
Household Effective Buying Income	5,775
Per Capita Total Local Taxes	143
Per Capita Taxes as Percent of Per Capita Income	6.1%
Percent of Unemployment	6.5%

TABLE 3

PUBLIC SCHOOL ENROLLMENT BY SCHOOLS, 1975

	TOTAL	NON-WHITE	PERCENT	WHITE	PERCENT
HIGH SCHOOLS					
Bridge	1554	758	48.8%	796	51.2%
Ives	2756	280	10.2%	2476	89.8%
Banbury	1461	974	66.7%	487	33.3%
JUNIOR HIGH SCHOOLS					
Margery Daw	928	477	51.4%	451	48.6%
Hiram Gordon	621	375	60.4%	246	39.6%
Elsie Marley	1034	249	24.1%	785	75.9%
Polly Flinders	934	97	10.4%	837	89.6%
William McTrimbletoe	723	14	1.9%	709	98.1%
Robert Barnes	687	451	65.6%	236	34.4%
John Watts	792	558	70.5%	234	29.5%
ELEMENTARY SCHOOLS					
L.J. Horner	367	358	97.5%	9	2.5%
P.P. Pumpkineater	808	433	53.6%	375	46.4%
Knaveofhearts	630	365	57.9%	265	42.1%
Jackandjill	786	595	75.7%	191	24.3%
H.D. Dock	651	309	47.5%	342	52.5%
M. Goose	660	5	.8%	655	99.2%
Lamb	719	39	5.4%	680	94.6%
L.M. Muffet	455	5	1.1%	450	98.9%
J. Sprat	457	10	2.2%	447	97.8%
Calicocat	772	60	7.8%	712	92.2%
T. Blindmice	666	6	.9%	660	99.1%
H.D. Diddle	621	221	35.6%	400	64.4%
O.K. Coke	762	265	34.8%	497	65.2%
Shaftoe	766	36	4.7%	730	95.3%
M.M. Contrary	339	0	0%	339	100%
Inashoe	662	150	22.7%	512	77.3%
S. Simon	616	347	56.3%	269	43.7%
Hubard	554	337	60.8%	217	39.2%
Sixpence	420	420	100%	0	0%
B.B.B. Sheep	441	441	100%	0	0%
L.B. Blue	241	0	0%	241	100%

SECTION II: Concern about Seventh Grade Arithmetic

Many of the Grandview high school mathematics teachers have recently complained that a large number of seventh-grade students are unable to add, subtract, multiply, and divide well enough to succeed in the regular curriculum. Since the teachers believe that mastery of simple computational skills is prerequisite to success in all walks of life, a joint teacher-administrator-parent committee has been formed to investigate various means of alleviating the problem. In a preliminary report submitted to the School Board, the committee proposed consideration of a remedial arithmetic program called AUTOMATH as a solution to the problem, and recommended that a thorough assessment of the problem, and this AUTOMATH program be conducted by the school system.

PROPOSED SOLUTION

In the AUTOMATH program, students work with a computer, through a terminal in the classroom. The computer automatically administers a series of drill-and-practice exercises in the basic computational skills (addition, subtraction, multiplication and division). Lessons range from elementary computation exercises involving single digit numbers to problems in applied arithmetic which require a high level of skill in decimal arithmetic.

AUTOMATH also has a set of computer actuated tape recordings which give vocal instructions concomitant with the printed instructions being typed out at the computer terminal. The computer automatically keeps tract of each student's progress and routes him to exercises which are appropriate to his current level of performance.

A Profile of Margery Daw Junior High School

Margery Daw Junior High School is the third largest of 7 junior high schools in Grandview. It was built in 1950, and is located on a large site in the North End neighborhood. Its attendance area includes part of the Bakersfield neighborhood and the entire neighborhood of North End. The building was renovated in 1970, and at present has one "portable" classroom in service.

Bakersfield, named for one of Gandview's leading industrialists, was originally developed by Mr. Ambrose Bakersfield's machine tool company, as a residence area for his factory employees, starting in the early 1900s. The modest but neatly kept houses are still inhabited largely by factory workers and their relatives, including both the skilled craftsmen and the semi-skilled workers. This area is fully integrated racially. The North End neighborhood contains mostly 3 or 6 family apartment buildings, plus some very small and inexpensive individual homes. Its inhabitants are predominantly black, but some Chicanos and other Spanish-speaking groups are represented there also. Most of residents are employed, and work in city as semi-skilled or unskilled workers.

According to the 1970 U. S. Census, 21,416 persons lived in the attendance area of Daw Junior High School at that time. About half (48.4%) were non-white.

The total enrollment at Daw Junior High last year year was 928, of whom 309 were in the 7th grade. Last year, 51% of the students in the 7th grade were non-white; this is somewhat above the percentage of non-whites across all the junior high schools in the city. Although the program at Daw is departmentalized, the homeroom groupings correspond roughly to the English

classes. There are 11 homerooms, ranging in size from 26 to 29 students each. The proportion of students enrolled in the occupational preparatory program at Daw is somewhat higher than the city average.

The school's facilities and equipment are felt by most of the teachers to be adequate. There are two sound movie projectors, three large television receivers, and a variety of slide and overhead projectors, etc. The mathematics group has a small collection of special teaching aids, including special films, models, and several new calculators. The average age of the textbooks used in the school, however, is 6.3 years. The enrollment last year was just below rated capacity, taking into account the extra classroom.

The principal, Ralph Babbage, feels that the school is generally functioning satisfactorily. There have been some minor problems of racial incidents and race tensions in the past, but last year was quite uneventful in this respect. There are very few serious discipline problems, and no evidence of any drug or alcohol problems among the students. However, absenteeism and occasional truancy have been a cause of some concern.

The faculty includes 33 persons, including a part-time nurse and a part-time social worker, each shared with 2 other junior high schools. Four of the faculty members are black, and the rest white. There are 11 men and 22 women faculty members. As of this fall, all of the teachers at Daw have had at least one year of teaching experience, and 12 of the staff have 20 or more years. Four of the teachers have master's degrees, and three others are actively pursuing such degrees now. For the entire faculty, the average number of semester credits beyond the highest degree is 10.7. There were five departures of faculty members from Daw last year, which is about the same as in previous years.

As part of their involvement in this study, the principal and a group of Daw School teachers have compiled the following tables, showing some information which provides more detailed description of last year's 7th grade classes. These tables show the average score for each class, and a measure of the spread or range of scores called the standard deviation, for several variables. The standard deviation measure of spread (labeled "S" in the table) is, very roughly, the number which is likely to contain one-third of the scores on each side of the average in most distributions. Some further notes are attached at the bottom of the table

Homeroom Teacher	% male	% white	Attendance* Average	Attendance* S	Otis IQ Average	Otis IQ S
Adams	.6	.4	3.9	0.5	98.6	11.0
Berkeley	.5	.5	3.7	0.8	99.9	14.4
Coleman	.5	.6	4.0	0.6	109.2	15.1
Davis	.6	.2	3.1	0.8	91.4	13.5
Edwards	.4	.4	3.1	0.9	89.7	11.3
Forrester	.5	.8	4.2	0.6	112.7	13.7
Gordon	.5	.6	3.9	0.6	104.4	14.6
Hunter	.6	.4	3.7	0.7	97.4	13.1
Jenkins	.6	.4	3.7	0.7	99.4	13.1
King	.5	.5	3.8	0.7	100.9	15.4
Lane	.5	.2	3.2	0.8	90.2	13.3
School Average	.5	.5	3.7	0.8	99.3	15.2

* The attendance score is based on a district-wide rating formula, in which a score of 5 is excellent, 4 is good, 3 is fair, 2 is poor, and 1 is unsatisfactory.